Dietmar Jannach, Lukas Lerche, Michael Jugovac

Item familiarity as a possible confounding factor in user-centric recommender systems evaluation

Recommender Systems, User Study, Satisfaction, Methodology

Summary. User studies play an important role in academic research in the field of recommender systems as they allow us to assess quality factors other than the predictive accuracy of the underlying algorithms. User satisfaction is one such factor that is often evaluated in laboratory settings and in many experimental designs one task of the participants is to assess the suitability of the system-generated recommendations. The effort required by the user to make such an assessment can, however, depend on the user's familiarity with the presented items and directly impact on the reported user satisfaction. In this paper, we report the results of a preliminary recommender systems user study using Mechanical Turk, which indicates that item familiarity is strongly correlated with overall satisfaction.

1. Introduction

Recommender systems are nowadays an integral part of many online services like shopping sites, Social Web platforms, or media streaming services. The main functionality of such systems is usually to filter and rank larger sets of information items in a personalized way and to reduce the information overload for the user (Jannach et al., 2010).

Measuring the success of a recommender system in practice can be done by determining business-related measures like click-through rates, conversion rates, or customer retention. Different variations of the system can be benchmarked with the help of A/B tests, in which alternative versions of the system are deployed for different user groups. In academic research settings, these types of system evaluation and comparison are usually not possible. Instead, academic studies are often based either on setups that rely on historical log data comprising, e.g., purchase transactions of a larger user community or on user studies that are conducted in a controlled environment. Using log data and evaluation measures like precision or recall from the Information Retrieval field can for example help to assess the predictive accuracy of different (machine learning) algorithms. User studies, on the other hand, can be designed to evaluate if the users actually found the system helpful or if they were satisfied with it as a whole.

Research in the field of recommender systems is nowadays dominated by a small set of well-defined offline experimental designs. User-centric evaluations, on the other hand, are often specifically designed for the given research problem. Only in recent years, more general frameworks for user-centric recommender system evaluation have been proposed, which should help to standardize the research methodology and thereby increase the reproducibility and comparability of research results (Pu et al., 2011, Knijnenburg et al., 2012).

1.1 Problem Statement

In typical laboratory study setups, users interact with slightly different versions of a recommender system. The goal is to understand the effects on the users when varying different aspects of the system, e.g., the algorithms that are used to generate the recommendation lists (Ekstrand et al., 2014) or the types of explanations that are presented (Gedikli et al., 2014). In the research designs of (Said et al., 2013) or (Ekstrand et al., 2014) for example, users were asked to assess the presented recommendations in different dimensions like novelty, usefulness, or diversity. Users either evaluated individual lists or made side-by-side comparisons of two different lists. Beside list-specific assessments, it is common in these studies that the participants are asked about their overall satisfaction with the system as well as their intention to re-use the system or recommend it to friends, as proposed, e.g., in the Technology Acceptance Model or in (Pu et al., 2011).

A problem when asking participants to evaluate the quality of recommendation lists is that the required effort for accomplishing this task depends on whether or not the participant is familiar with the presented items. In (Said et al., 2013), the participants were explicitly asked for each recommended movie if they knew it. If not, additional information on the actors or the plot was made available to the users in order to assess if they would like to watch it. In their quantitative analysis, the authors looked at how many unfamiliar or unknown items were presented by each algorithm and if the users were interested in them. However, they did not analyze if the level of item familiarity had an influence on other aspects like the intention to reuse. In (Ekstrand et al., 2014), "Novelty" was modeled with multiple, slightly different questions, e.g., asking the user which of the presented two lists contained more "familiar", "pleasantly surprising", or "unexpected" items. Their analysis revealed that "Novelty" had a strong negative impact on their "Satisfaction" construct, which was partially based on the user's intention to reuse the system. The relationship to "Effort", however, was not examined in their research.
The decision-making effort can play a major role in the acceptance of a recommender system (Chen and Pu, 2009). In (Bettman et al., 1990) it has been shown that users were satisfied with somewhat inaccurate decisions if they led to a reduction in their perceived effort, since the latter is more directly tangible. Not only the objective effort, quantified, e.g., by the number or time of interactions (McCarthy et al., 2005), should be measured but also the perceived effort of the users (Payne et al., 1993).

1.2 Outline

One main hypothesis of the research presented in this paper is that item popularity and familiarity can have a strong impact on the user's quality perception of the overall system, partially because the assessment of recommendation lists containing fewer known items requires more cognitive effort by the user. In the paper, we report the results of a user study that is similar to the ones conducted by (Said et al., 2013), (Ekstrand et al., 2014), and (Cremonesi et al., 2013) as we let users assess recommendations generated by different algorithms in several quality dimensions. In our evaluation, we, however, also specifically look at item familiarity as a possibly relevant factor for the participants' overall assessment of the system. Furthermore, we briefly discuss how we dealt with the challenges of conducting user studies on a crowdsourcing platform like Mechanical Turk.

2. Study Design

The goal of our study was to assess how the recommendations generated by different algorithms are perceived by users and if item familiarity plays a role in the user's perception of the system. An additional aim was to determine whether or not "objective" measures used in the literature, e.g., to assess the prediction accuracy or list diversity, correlate with the reported perceptions of the participants. We used the following study design.

2.1 Tasks for the Participants

The study setup is generally quite similar to the recent studies mentioned above. The participants in our study had to accomplish three tasks using an online web application that was created for the study.

(1) As a first step, the users had to provide ratings for at least 15 movies they had seen in the past using a 5-star scale with half-star increments. The movies to be rated could be picked from a randomly generated list; alternatively, the users could use a search field to retrieve movies they know.

(2) In the second phase, the users were presented with 10 movie recommendations and had to individually rate the movies. We provided detailed information for the movies (e.g., actors, genres, plot synopsis, trailers) so that the users could provide ratings even if they did not know the movies. In any case, the participants had to explicitly state whether or not they knew this movie.

(3) In the third phase, we presented the list of rated and recommended movies on one page and asked the users to answer a set of questions regarding, e.g., whether they found the set of recommended items to be diverse and/or a good match for the given preferences.

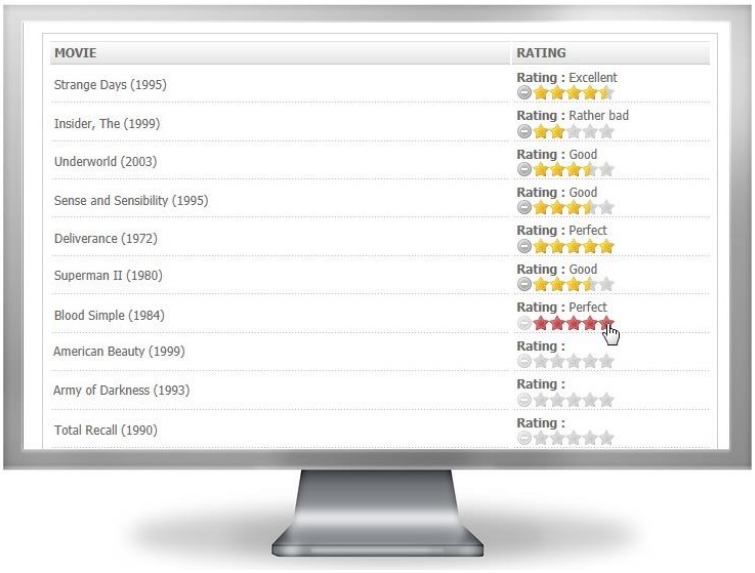

Figure 1: Step one - preference elicitation (at least 15 movies to be rated)

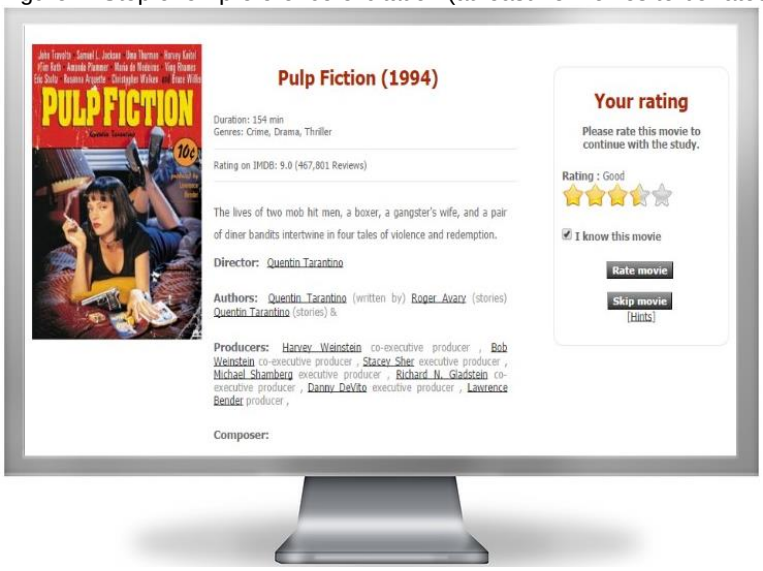

Figure 2: Step two - rating 10 recommended movies individually

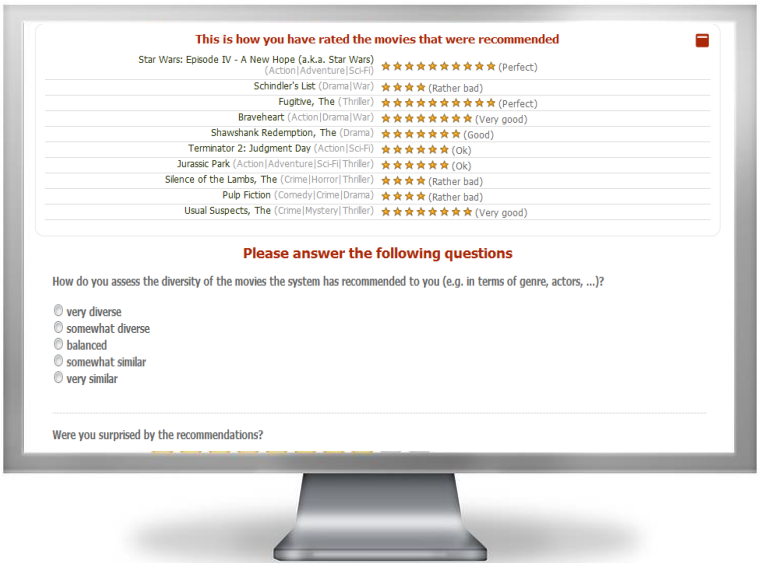

Figure 3: Step three - assessing the list in different dimensions

The following set of questions was asked at the end of the survey. Except for question 1, the answers were given on five-point Likert scale.

1) How well do the recommended movies match your general preferences?
   Answers on a rating scale from 1 to 10.
2) Would you consider yourself a movie enthusiast?
   Answers from "definitely not" to "definitely yes".
3) How do you assess the diversity of the movies the system has recommended to you (e.g. in terms of genre, actors, ...)?
   Answers from "very similar" to "very diverse".
4) Were you surprised by the recommendations?
   Answers from "not at all" to "everything was surprising".
5) Can you see a logical connection between the preferences and the recommendations?
   Answers from "not at all" to "everything was totally clear".
6) Would you recommend the system to a friend?
   Answers from "definitely not" to "definitely yes".
7) Was the system easy to use?
   Answers from "too complicated" to "very easy".
8) Would you use this or a similar system again?
   Answers from "definitely not" to "definitely yes".

2.2 Data set and algorithms

As an underlying rating database we used a subset of the MovieLens10M dataset for which we crawled content information from the IMDb website. Since updating the models of our algorithms with the new ratings of each participant can be computationally complex, we used a subsample of 400.000 ratings.

To generate personalized recommendations for the individual participants, we temporarily added the ratings provided by the users to the dataset and re-computed the models. In order to be able to compare recommendation techniques from different families, we used the following algorithms in our evaluation; see (Jannach et al., 2013) for more details.

- PopRank: A method that simply returns the most popular movies in the dataset, i.e., those movies for which the most ratings exist.
- SlopeOne: A comparably simple collaborative filtering (CF) method (Lemire and Maclachlan, 2005) that is about as accurate as user-based nearest neighbor methods on small datasets but can be computed faster.
- Funk-SVD: A highly accurate CF method based on matrix factorization.
- Bayesian Personalized Ranking (BPR): A recent learning-to-rank method designed for implicit feedback scenarios.
- Content-based Filtering (CB): A content-based method that recommends items that are similar to those that the individual user has liked in the past.

The assignment of algorithms to the participant was based on a random selection. The selection process was, however, biased in a way that each algorithm was used by about the same number of participants. Since the time needed to re-compute the models can vary across the algorithms, we used the running time of the slowest technique (Funk-SVD) as a reference time and delayed the response of all other algorithms accordingly to avoid any effects caused by the response times of the algorithms.

2.3 Participants

We used the Mechanical Turk crowdsourcing platform to recruit participants (“Turkers”) for the study. The Turkers were paid $1.50 and had to do a qualification test before they could participate. The qualification test was necessary because pilot tests revealed that many participants were unreliable and did not carefully rate the movies and answer the questions. In addition to the qualification test, we included a number of "attention checks" and excluded all participants who failed these. The following checks were implemented in the survey system:

- Non-existent movies: About every 5th movie in the list of movies to be rated in the first phase was made up. Since the participants were supposed to rate movies they have seen, a rating for a non-existing movie is a clear indicator of a lack of attention.

- Rating time: If the participants were "too fast" in the rating phase and needed less than a second for a rating, we considered this as a sign that their responses will not be reliable and excluded them.
- Questionnaire: In the questionnaire, we included two attention checks. First, we asked one question twice but with a reversed order of the answer options. Second, we included a question that contained an instruction on which answer had to be chosen. Participants who answered these control questions inconsistently or not according to the instruction were removed from the survey.

Overall, from the 175 participants, only 96 successfully completed the survey, which we see as an indicator that our attention checks were well-suited to identify non-reliable participants. The observations for each of the five algorithms are therefore based on about 20 participants.

3. Analysis of Observations

In the following sections, we will first report our observations on how the recommended items were perceived by the different participant groups and we will then have a closer look at correlations between the observations.

3.1 Perceived and Measured Accuracy

Being able to recommend items that match the preferences of a user is a desirable and central characteristic of a recommender system (even though being accurate might not be enough in some application domains). The first aspect that we analyze is therefore whether the recommendations of some algorithms match the tastes and preferences of the users better than others.

Perceived Accuracy

Figure 4 shows how the participants assessed the capability of each algorithm to recommend movies that match their general preferences. Quite surprisingly, the **non-personalized** method PopRank obtained the best feedback by the study participants. BPR, an algorithm that also has a strong bias to recommend popular movies, is ranked second, followed by the content-based technique. The techniques Funk-SVD and SlopeOne, which are designed for the rating prediction task, received rather mediocre feedback. They are significantly different ($p < 0.05$) to both PopRank and BPR; note that we use pairwise t-tests with Bonferroni correction throughout the paper for significance testing. Also, SlopeOne is significantly different to the content-based technique. An interesting aspect to note here is that the recommendations produced by the content-based method are comparably well-accepted by the users. Users seem to recognize that the content-based technique recommends items that are similar to what they have liked in the past and appreciate those recommendations even though they are not blockbusters and often not known to them as will be seen later in Table 1.

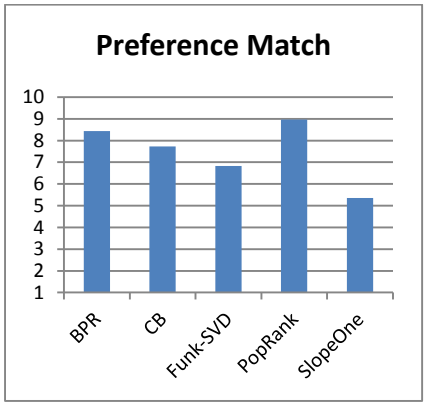

Figure 4: Answers to questionnaire item 1: Match of preferences.

Ratings provided for recommended movies

A different way of measuring the recommendation accuracy is to look at the ratings that were provided by the participants for the recommended movies. The results are shown in Figure 5 and we can observe that the ranking of the algorithms

equivalent to Figure 4, i.e., the participants were consistent in their behavior. PopRank is significantly different ($p < 0.05$) to the content-based technique, Funk-SVD and SlopeOne; SlopeOne also differs from BPR.

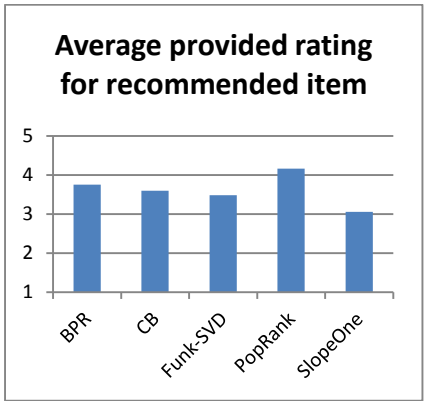

Figure 5: Average rating provided for the recommended movies

Measured Accuracy – RMSE and Precision (Offline experiment)

Our next aim was to quantify the prediction accuracy of the different algorithms in offline and online settings using the standard accuracy measures RMSE and precision.

Table 1: Accuracy measures; RMSE is not applicable for PopRank and BPR.

		PopRank	BPR	CB	FunkSVD	SlopeOne
1	Precision (Movielens)	0.62	0.64	0.62	0.69	0.68
2	Precision (Survey, all)	0.74	0.57	0.50	0.47	0.23
3	Precision (Survey, only 'not seen')	0.42	0.33	0.28	0.22	0.14
4	Precision (Survey, only 'seen')	0.76	0.59	0.58	0.70	0.48
5	RMSE (Movielens)	n.a.	n.a.	1.92	1.65	1.72
6	RMSE (Survey, all)	n.a.	n.a.	2.19	3.46	4.03
7	RMSE (Survey, only 'notseen')	n.a.	n.a.	2.90	4.55	4.45
8	RMSE (Survey, only 'seen')	n.a.	n.a.	1.87	1.99	2.59
9	Percentage of 'seen' movies	0.94	0.94	0.74	0.52	0.27

We report the accuracy results obtained through a standard cross-validation procedure using the underlying MovieLens dataset in the rows 1 (Precision) and 5 (RMSE) Table 1. The results are in line with what is typically reported in the literature. The personalized methods are better than the non-personalized baseline with respect to precision and Funk-SVD outperforms the other algorithms slightly. RMSE measurements could only be made for the rating prediction methods and we again see the common pattern that the matrix factorization method Funk-SVD performs better than the others.

Measured Accuracy – RMSE and Precision (Survey data)

Next, we compared the predictions and recommendations made by the algorithms with the explicit rating feedback provided by the users in the second phase of the survey. The results are shown in rows 2-4 (Precision) and 6-8 (RMSE) of Table 1. We furthermore split each measurement into two subgroups: one comprises the feedback on movies that the participants already knew ('seen') and one comprises movies that were unknown to the participant before the experiment ('not seen'). The percentage of known movies in the recommendations for each algorithm is given in row 9. The following observations can be made.

- **Popularity aspects:** Our first observation is that the movies recommended by PopRank and BPR are in fact popular and nearly all of them are known by the participants. The content-based method by design recommends comparably well-received movies that many users are familiar with. The rating prediction methods Funk-SVD and SlopeOne, on the other hand, seem to partially recommend niche (obscure) movies, a phenomenon that was also reported in the study by (Ekstrand et al., 2014).

- **RMSE:** When considering only the movies the user already knew (row 8), we can observe that (a) the content-based method and Funk-SVD produced the most accurate predictions and that (b) the error values in these cases are not too different from those that were obtained in the offline evaluation.
  However, the ratings provided by the participants for unfamiliar movies (row 7) are much harder to predict. Especially the performance of Funk-SVD and SlopeOne decreases strongly because the predictions are much higher than the provided ratings. We can speculate that (a) the algorithms either largely overestimate the preference of the participants for the recommended movies or (b) the participants were conservative in their ratings because they had not actually seen the movies. The performance of the content-based method also decreases for unfamiliar movies, but not to such a large extent as can be observed for the other methods.
- **Precision:** The observations for precision follow a similar trend as for the RMSE and the results obtained in the offline evaluation roughly correspond to what is measured in the survey for the known movies (row 4), with the exception of SlopeOne. For the movies unknown to the participants, the precision values strongly decreased (row 3). Note, however, that for BPR and PopRank the fraction of unknown movies is very small and therefore their results are only based on few ratings.

Overall, the popularity of the items and correspondingly the fraction of movies that were already familiar to the participants seem to be directly related with the perceived capability of the system to match the users' preferences. Algorithms like Funk-SVD and SlopeOne furthermore exhibit some tendency to recommend niche items. For these algorithms the average predicted item rating in the MovieLens dataset is actually higher than for the others, i.e., they generally recommend high-quality movies as shown in Table 2. The average rating provided in the survey is, however, lower for these methods as shown in Figure 5.

Table 2: Average MovieLens rating of recommended movies (on the original scale from 1 to 10)

Algorithm	Avg. ML rating
BPR	7.61
CB	7.36
Funk-SVD	8.44
PopRank	8.28
SlopeOne	8.46

Summary of observations:

- The participants felt that the non-personalized method that recommended popular movies was best at matching their personal preferences.
- The participants gave higher ratings for popular movies (which they most likely knew) than to movies that actually had a higher average community rating in the MovieLens dataset.
- The participants gave lower ratings to movies which they did not know. Comparing predicted and actually provided ratings as a means to assess the predictive accuracy of different algorithms in such user studies seems to be unreliable, as the observed results largely depend on whether the participants know the movies or not.

3.2 Diversity, Surprise, and Transparency

In this section, we will discuss other often mentioned possible quality factors for recommender systems: diversity, surprise and transparency.

Questionnaire results

The results obtained in the questionnaire in the last phase of the experiment are shown in Figure 6.

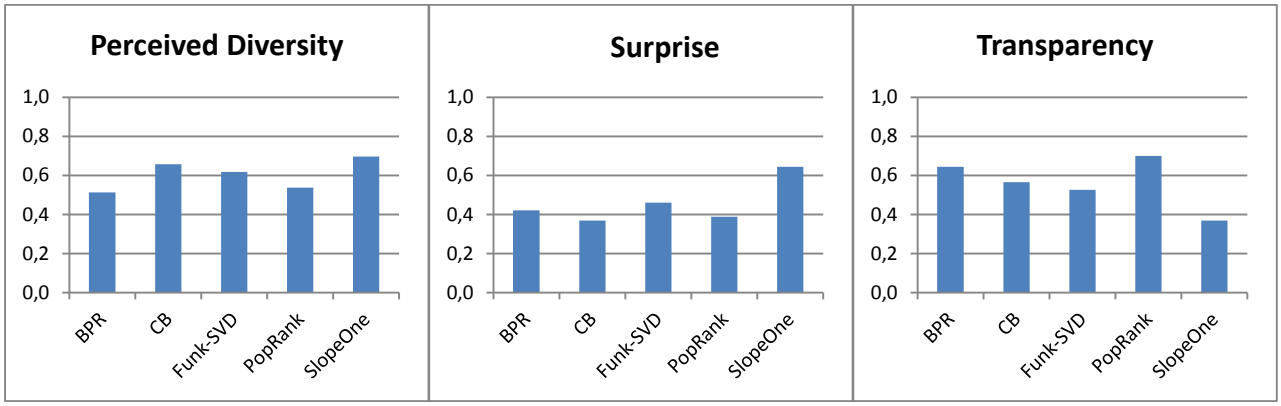

Figure 6: Perceived diversity, surprise, and transparency

- **Diversity** The level of perceived diversity in terms of the movie genres for the different algorithms is shown in Figure 6 (a) and based on the survey question "How do you assess the diversity of the movies the system has recommended to you (e.g. in terms of genre, actors, ...)?" from "very similar" to "very diverse". The popularity-based methods PopRank and BPR achieve values which were considered by the participants to be a good balance of genres (0.5 = "balanced"). All other algorithms tend to recommend items that are perceived as slightly more diverse. The recommendations of SlopeOne, which are mostly based on high average item ratings and not on genre preferences, led to the highest diversity impression.

- **Surprise:** When asking the participants if they "were surprised by the recommendations" ("not at all" to "everything was surprising"), they found only "very few" to "some surprises" in the recommendations generated by all algorithms except for SlopeOne (see Figure 6 (b)). The level of serendipity that can be achieved through these methods and the corresponding chances of discovering unexpected but relevant movies are thus comparably low. SlopeOne is significantly different to BPR, PopRank and the content-based technique ($p < 0.05$) and proposes movies that were on average "quite surprising" for the study participants.

- **Transparency:** To assess the transparency of the recommendations, we asked if the users could find "a logical connection between the preferences and the recommendations" ("not at all" to "everything was totally clear"). We can observe a phenomenon comparable to the preference match of the algorithm. Even though PopRank is non-personalized, the participants felt that the method's rationale of recommending item was on average "mostly obvious" (Figure 6 (c)). Again, recommending popular items seems to be the best strategy. For the content-based method and Funk-SVD, the participants on average stated that "some logic can be seen"; the niche recommendations of SlopeOne, finally, were "mostly unclear" to the participants and significantly different ($p < 0.05$) to BPR and PopRank.

Subjective and objective diversity

Similar to the work of (Cremonesi et al., 2013) we compared the perceived diversity impression with an objective measure. In our case, we computed the diversity of the recommendation lists of the different algorithms using the inverse of the intra-list similarity measure (ILS, i.e., the average pairwise similarity). The measure was determined using TF-IDF representations of the movie descriptions and cosine similarity as a distance measure.

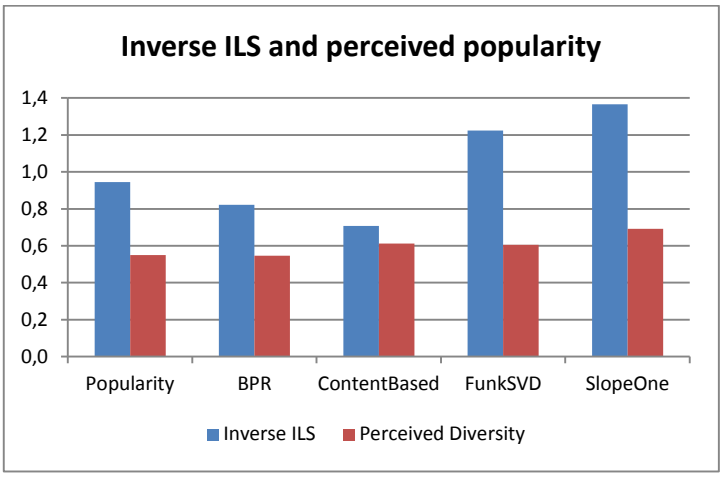

Figure 7: Objective and subjective diversity

We show the results obtained for the perceived diversity and the inverse ILS (intra-list diversity) in Figure 7. When looking at the inverse ILS, the results are not very surprising: the content-based method as expected leads to the lowest diversity values because the algorithm by design recommends items with similar TF-IDF item representations. Funk-SVD and SlopeOne are very different from the other techniques and seem to recommend highly diverse sets of items.
Regarding the user perception, however, there seems to be no difference between Funk-SVD and the content-based method. The popularity-based method, on the other hand, leads to a comparably low diversity perception despite its relatively high inverse ILS value.

Summary of observations:

- The objective Intra-List-Similarity measure did not always correlate with the user-perceived diversity in our experiment and it is unclear if the ILS measure is suitable to assess the diversity of recommendations in offline experiments.
- The participants found recommendations slightly more diverse when they contained movies which they did not know.
- The participants found the reasoning logic of the algorithms more transparent when the system recommended popular movies known to the participants or when they were based on a content-based algorithm. The stated level of surprise for these algorithms was consequently lower for these algorithms.

3.3 User acceptance

The final set of questions in the questionnaire was focused on usability and the acceptance of the system as a whole. The statistics obtained from the questionnaire are shown in Figure 8.

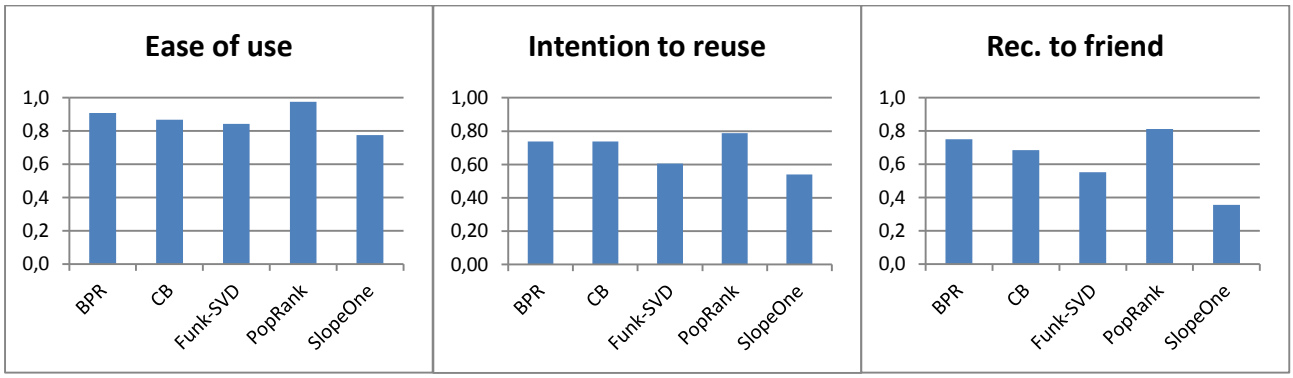

Figure 8: Ease of use, intention to reuse, recommendation to a friend

Questionnaire results

- **Ease of use:** The answers regarding the ease of use ("Was the system easy to use?" from "too complicated" to "very easy") was consistently high for all algorithms. Even though the results for Funk-SVD and SlopeOne are lower than for

the other algorithms (just "easy"), their differences are only significantly different to PopRank ($p < 0.05$). Since these algorithms tend to recommend more diverse and niche items, some users might struggle to judge unfamiliar items and therefore perceive the system as more difficult to use.

- **Intention to Reuse:** Asking "Would you use this or a similar system again?" ("definitely not" to definitely yes") leads to significant differences in the user's answers. While users of BPR, CB and PopRank are inclined to come back to the system ("probably yes"), users from the latter tend to be unsure ("maybe"). The intention to reuse for the content-based method is not different from BPR and PopRank, even though its perceived accuracy was lower.

- **Recommendation to a friend:** The feedback on the user's intention to "recommend the system to a friend" ("definitely not" to "definitely yes") exhibits a similar but stronger trend. In comparison, most users would "rather not" recommend a SlopeOne-based system to friends while Funk-SVD would "maybe" be recommended. Participants who were served by one of the other algorithms are mostly inclined to tell others ("probably yes"). The answers for SlopeOne are significantly different to those for BPR, PopRank and the content-based technique ($p < 0.05$). Also, Funk-SVD is significantly different to PopRank.

Objective user effort

Since the effort required by the user to interact with the system can have an impact on the user acceptance, we measured the time needed by the users to rate the recommended movies. On average, the participants needed 75 seconds for PopRank, 93 for BPR, 163 for CB, 211 for Funk-SVD and 316 for SlopeOne. While the differences between PopRank and BPR, as well as between CB and Funk-SVD are not significant, all others are ($p < 0.05$).
Since some algorithms recommend more unfamiliar movies, their users most probably needed more time to study the provided additional information about the movie on the survey page. The higher effort did, however, not impact on the reported ease of use, for which no differences were observed.

Summary of observations

- The participants are more inclined to reuse a system or recommend it to a friend when it recommends popular items (which the users know and like) or when the items are similar to those that the user has positively rated.
- The required user effort for completing the survey had no impact on the perceived use of the system.

3.4 Observed Correlations

The results presented in the previous section indicate that some of the answers provided by the users and different objective measurements are correlated, e.g., that users assigned higher ratings to movies that they already knew. In the following section, we will report how strong some of the variables are correlated using Spearman's rank correlation coefficient (ρ). Given our study design and our basic correlation analysis, we can, however, make no claims about causation. The reported observations should thus serve as a basis for future studies in which the corresponding hypotheses are developed and tested through an appropriate experimental design and statistical methods that are suited to identify causal relationships.

Perceived accuracy

Accuracy is often the main goal in research in recommender systems. In fact, our data shows that higher levels of perceived accuracy are positively correlated with higher perceived transparency ($\rho = 0.67$), the intention to reuse the system ($\rho = 0.55$) and the intention to recommend the system to a friend ($\rho = 0.75$). These three dimensions are an indicator for the users' satisfaction.
This seems like a good sign at first glance. Remember, however, that the highest levels of perceived accuracy in our study were obtained when a non-personalized method was used or when mostly popular items were recommended.

Item familiarity

The observations reported so far suggest that item familiarity, i.e., the extent to which participants already knew the recommended movies, is strongly connected with other variables, e.g., how the users rated the movies (Section 3.1).

The more specific correlations are as follows:

- The higher the percentage of known items in the recommendations, the higher is the perceived accuracy (Figure 1), ρ = 0.68, and the higher is the average rating provided by the participants (ρ = 0.60).
- When there are more known items presented to the user, also the perceived transparency is higher (ρ = 0.53).
- Finally, we can observe that higher item familiarity is positively correlated with the intention to reuse (ρ = 0.40) and the tendency to recommend the system to a friend (ρ = 0.55).

Given only these correlations, we cannot know if item familiarity is indeed the reason for higher satisfaction or if there are other latent factors causing the correlations. Nonetheless, we see the strength of the correlations as a sign that further investigations are required. In particular, we have to validate that item familiarity is not a confounding factor in user studies in which the participants have to rate movies they do not know or assess recommendations lists containing varying levels of known items as in (Ekstrand et al., 2014).

In (Ekstrand et al., 2014), the construct "Novelty" was found to have a strong negative impact on user satisfaction. Since item familiarity is one of several questionnaire items to assess "Novelty" in their work (including surprise), a more detailed analysis is required to isolate the role of item familiarity alone.

Surprise and perceived accuracy (serendipity)

The questionnaire item on surprise was added to find out if the algorithms are capable of producing serendipitous recommendations, which are both surprising and useful (accurate). Overall, however, the level of surprise is consistently negatively correlated with the perceived accuracy across all recommendation techniques (ρ = -0.54). Again, in the study of (Ekstrand et al., 2014), surprise is a part of the "Novelty" construct, which negatively impacts satisfaction.

Diversity and user satisfaction

Diversity was found to have a slight positive impact on user satisfaction in (Ekstrand et al., 2014). In Section 3.2, the data indicated that perceived diversity and an objective measure like ILS are not strongly connected, which was confirmed by the correlation analysis (ρ = 0.31). Furthermore, the correlation of perceived diversity with any of the factors related to user satisfaction was modest at best. Therefore, we see no strong indication in our study results that diversity actually helps to increase the overall satisfaction with the system.

User effort and user satisfaction

We measured the objective user effort in terms of the time needed to rate the recommended movies and we have seen in Section 3.3 that users needed less time when they had to rate popular movies or movies they already know (ρ = -0.62). User effort is also correlated negatively with the perceived accuracy (ρ = 0.52) and slightly with the average rating of the recommendations (-0.34). At the same time there also exists a negative correlation between the perceived ease of use with the perceived accuracy (ρ =0.39) and the intention to reuse the system (ρ = 0.42).

4. Discussion and Outlook

Summary

In our user study, recommending the most popular items to everyone, i.e., the movies that most participants were familiar with, was the strategy that led to the highest acceptance and perceived recommendation quality. In practice, however, recommending popular items might be of little value for the customer and the business (Jannach and Hegelich, 2009). At the same time, we could observe that users gave particularly low ratings to movies which they did not know already. The offline determined accuracy measures did not correspond to the prediction accuracy as measured in the user study.

Overall, our Mechanical Turk experiment leads us to the hypothesis that item familiarity can be a confounding factor in user studies of the type presented in this paper and in previous studies of a similar type. Like us, (Ekstrand et al., 2014) observed that some algorithms like Funk-SVD exhibit a tendency to recommend quite obscure niche items. In their study, they therefore set a threshold on the minimum popularity of the recommended items.

Limitations

The results presented in this paper are based on a study comprising 96 participants, i.e., there were only about 20 participants in each group. Furthermore, even though we did not include unreliable participants and the remaining ones were unbiased and working carefully, they were still monetarily motivated.

Another possible limitation is that in our survey we directly asked questions related to the quality dimensions of interest. However, asking the user directly might be too technical or lead to a misinterpretation of the questions' true meaning. Using more than one question per dimension, as in (Ekstrand et al., 2014), might alleviate this effect but in turn introduces other problems, e.g., how to weight the different aspects or how to ensure that the construct really reflects what has been asked. If multiple questions are asked, it is therefore essential to verify their internal consistency as for example done in (Pu et al., 2011).

Conclusions and future work

As an general result, we argue that more experiments are required to better understand the role of item familiarity in recommender systems user studies as otherwise the reliability of the results can be limited.

Since recommendation lists with high item familiarity were well perceived by the users, one future research approach could be to create recommendation lists that contain both familiar items – so as to increase the user's trust in the system – and new items to help the user discover new and relevant items. To the best of our knowledge, no research has carried out so far in that direction.

5. References

Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems - An introduction, Cambridge University Press, 2010.

Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems, Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11), pp. 157-164, 2011

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction, 22(4-5), pp. 441-504, 2012

Chen L., and Pu, P.: Interaction design guidelines on critiquing-based recommender systems, User Modeling and User-Adapted Interaction 19(3), pp. 167-206, 2009

Gedikli, F., Jannach, D., Ge, M.: How should I explain? A comparison of different explanation types for recommender systems, International Journal of Human Computer Studies, 72(4), pp. 367-382, 2014

Ekstrand, M.D., Maxwell Harper, F., Willemsen, M.C., Konstan, J.A.: User perception of differences in recommender algorithms, Proceedings of the Eight ACM Conference on Recommender Systems (RecSys '14), pp. 161-168, 2014

Said, A., Fields, B., Jain, B. J., Albayrak, S.: User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm, Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '14), pp. 1399-1408, 2014

Cremonesi, P., Garzotto, F., Turrin, R.: User-centric vs. system-centric evaluation of recommender systems, Proceedings INTERACT 2013, pp. 334-351, 2013.

Jannach, D., Lerche, L., Gedikli, G., Bonnin, G.: What recommenders recommend - An analysis of accuracy, popularity, and sales diversity effects, 21st International Conference on User Modeling, Adaptation and Personalization (UMAP '13), pp. 25-37, 2013

Bettman, J.R., Johnson, E.J., Payne, J.W.: A componential analysis of cognitive effort in choice, Organizational Behavior and Human Decision Processes, 45(1), 111-139, 1990

Payne, J.W., Bettman, J.R., Johnson, E.J.: The Adaptive Decision Maker, Cambridge University Press, 1993

McCarthy, K., McGinty, L., Smyth, B., Reilly J.: On the evaluation of dynamic critiquing: A large-scale user study, Twentieth National Conference on Artificial Intelligence (AAAI '05), pp. 535-540, 2005

Jannach, D., Hegelich, K.: A Case Study on the effectiveness of recommendations in the mobile internet, Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09), pp. 205-208, 2009