Expectation, Perception, and Accuracy in News Recommender Systems: Understanding the Relationships of User Evaluation Criteria Using Direct Feedback

Poornima Belavadi, Laura Burbach, Stefan Ahlers, Martina Ziefle, and André Calero Valdez(✉)

Human-Computer Interaction Center, RWTH Aachen University, Campus Boulevard 57, 52076 Aachen, Germany
{belavadi,burbach,ahlers,ziefle,calero-valdez}@comm.rwth-aachen.de

Abstract. Typically, a user-focused approach of evaluation of recommender systems requires the users to recollect their experiences, exposing study results to memory biases. In this paper, we describe a study conducted to test a framework, that allows recommender systems to be used and evaluated simultaneously. In this study, we asked 140 participants about their expected, perceived, and actual quality of the recommendations. We compare the performance of two recommender systems. The singular value decomposition recommendation system was able to correctly predict more than half of all evaluations and performed better than participants expected. However, users were more satisfied with the suggestions of the user-based collaborative filtering recommendation system. Our approach allows to compare actual item ratings, expected quality, and perceived quality of recommendations. Serendipity was found to be an important influencing factor for better item ratings by users. Participants rated both recommendation systems better when they perceived higher quality.

Keywords: Recommender systems · Live evaluation · User studies · Recommendation accuracy · News recommendation · Qualtrics

1 Introduction

Recommender systems use different algorithms to predict what a user likes. By utilizing user data, recommender systems draw conclusions about the preferences and interests of the user. With multinational companies using recommender systems (e.g., Amazon, Netflix etc.) to build revenue, research on improving recommender systems has been gaining interest. A good portion of which focuses on evaluation of recommender systems. The evaluation is done either by mathematically determining the accuracy of the algorithm or by conducting user

C. Stephanidis et al. (Eds.): HCII 2021, LNCS 13094, pp. 179–197, 2021.
https://doi.org/10.1007/978-3-030-90238-4_14

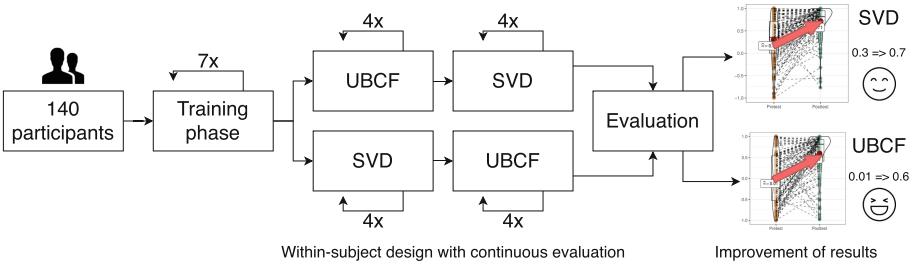

Fig. 1. User study showing the experimental design and results.

studies. Even in user studies, the recommender systems were mostly analyzed offline according to statistical criteria, such as run-time efficiency of the algorithms used or the accuracy of the recommendations. Only some studies focus on the attitudes of users of recommender systems and carried out online evaluations [18,34]. In these studies, participants were asked about their experiences with a recommender system retrospectively, but the studies did not check whether the reported experiences and opinions corresponded with actual recommendation quality.

In this paper, we describe a user study to test a framework that allows the users to use and evaluate two recommender systems (see Fig. 1). We asked the users to give feedback about the quality of the recommendations directly after exposure. The recommender system used this feedback and the data was used to re-train the algorithm instantly. We tested the framework on the evaluation of two *recommender systems for news articles*. The collected data enabled us to evaluate the recommender systems according to both statistical and user-centric criteria. We examine and compare the *quality of the recommended news items expected by the respondents*, the *perceived quality of the recommendations*, and the *actual item rating* while re-training the recommender system during use.

2 Related Work

Today, various recommender systems are being developed and used in many different application contexts such as *entertainment* [22,39], *online shopping* [7,41], *e-health* [9,12,47] and *social networks* [13,49], creating a need for better recommender systems. This has motivated research on the evaluation of recommender systems. Evaluating and developing recommender systems poses challenges. Following, we discuss some of the different types of recommender systems and challenges faced during their development.

2.1 Types of Recommender Systems

Most commonly, recommender systems are divided into three classes: *content-based recommender systems*, *collaborative filtering* and *hybrid recommender systems* [1,46]. Most recommender systems combine sub-types from different classes

to compensate for the weaknesses of the individual systems and benefit from the respective strengths [45].

Content-Based Recommender Systems. Content-based recommender systems give recommendations based on the attributes of items that have been evaluated in the past. This involves comparing attributes of the evaluated items with the attributes of items that have not yet been evaluated. If the attributes of a new item correspond to the attributes of items that the user rated well, the new item is recommended [45].

Collaborative Filtering. The most widespread recommender technique being used is collaborative filtering [2,8]. Here, the ratings of users are sampled to create a user-item rating matrix. In this matrix, each column corresponds to an item that has been rated and each row corresponds to a user who assigned the rating. The value at the intersection of row and column is the rating the user assigned to the corresponding item [38]. The absence of a value represents that the user has not yet given a rating for this item [38]. The recommender system predicts these values, to identify items that could be of interest to a user by comparing their past ratings with the ratings of other users [40,48]. Collaborative filtering can be divided into two classes: *memory-based* and *model-based filtering* [8,23,42].

Memory-Based Collaborative Filtering. Memory-based collaborative filtering provides recommendations by generating either *user-based* or *item-based predictions* for items by searching for similarities in the user-item rating matrix [23]: *User-based collaborative filtering (UBCF)* compares the similarity between users. Two users are similar if the ratings of the items they both rated are similar. To generate recommendations for all the items that have not yet been rated by the user, the ratings of these items by similar users are compared. The predictions of the items correspond to the weighted average ratings of the other users. The greater the similarity between the users, the more weight is given to the rating [2,23]. *Item-based collaborative filtering* compares the similarity between items and not between users. The system calculates the similarity between all item pairs based on their ratings. To generate new recommendations for a user, the system compares the similarity between their rated items and the items that have not yet been rated. The items that have not yet been rated and show the greatest similarities are recommended to the user [23].

Model-Based Collaborative Filtering. Model-based collaborative filtering uses various techniques to calculate how a user might rate an item. This is done using machine learning, but also statistical methods such as *Bayes networks*, *cluster analysis*, *matrix factorization* [40] or *Singular Value Decomposition (SVD)*, which is a method from linear algebra that is commonly used as a method for reducing dimensionality [50]. Compared to memory-based collaborative filtering, model-based collaborative filtering scales better and takes less time to calculate the recommendation, but the development of the underlying model is more costly [2].

For our study, we use UBCF as the representative for memory-based and we use SVD as the representative for model-based techniques.

2.2 Challenges in Developing Recommender Systems

Developers of recommender system face a multitude of challenges: As recommender systems operate on large pools of data, an overlap between two users may be very small or non-existent. In addition, the distribution of ratings among the individual items can be extremely unequal. A recommender system must take this lack of data (*data sparsity*) into account [21,30]. The problem of data sparsity is especially relevant for the so-called *cold-start problem*. To classify users and provide them with appropriate recommendations, recommender systems need information on them such as their ratings. For new users who have not yet submitted ratings, this information is missing (*cold-start problem - new user)*, hence, the recommender system cannot make accurate recommendations to these users [20,43,45]. New items cause problems in a similar way (*cold-start problem - new item)*. If recommender systems only consider the users' ratings and not the attributes of the items, they will never recommend items that have not yet been rated [20,43,45].

Another problem faced especially by content-based recommender systems is the problem of *overspecialization*. As they are solely based on the users' previous ratings, they only provide recommendations for new items that are similar to the ones the user has previously rated [46]. Some of these might be inappropriate to recommend (e.g., recommending a washing machine after the user has already purchased a different one).

A recommender system should also make unexpected recommendations to users that may prove to be appropriate (*serendipity*). To make frequent suitable recommendations, the system usually recommends items that are currently very popular with other users or are often rated highly. This is problematic because these items are often found by users without the help of a recommender system. Hence, a good list of recommendations should also contain less obvious items that the user would probably not find without the recommendations of the system. Balancing the accuracy and variety of recommendations is a central challenge for recommender systems [30,31,46].

Recommending News Articles. Recommending news articles poses further challenges [24,29]. Recommender systems must analyze and classify a large number of articles in a very short frame of time. This is further complicated by variations in the structure of different article types [29,51]. For no other item *topicality* plays such a decisive role. Articles that are interesting today can be uninteresting tomorrow. One requirement for recommender systems is therefore not to recommend articles that are no longer up-to-date [29,51].

To recommend suitable news articles, detailed *user profiles* must be created. The recommender system should automatically record the articles that the user has read, while preserving the privacy of the users [29].

There has been only sparse research on recommender systems for news articles (e.g., [6,11,17,26]). Beyond recommending news articles in real time [3,4,25,32], we connect real-time recommendation with a simultaneous evaluation of the recommender system.

2.3 Offline- vs. Online Evaluation

The evaluation of recommender systems is carried out in either of two ways: online or offline. For offline evaluation, data is first collected or simulated and the recommender system is then tested in a system-centered manner. By contrast, in online evaluation, real test users evaluate a prototype or a productively used recommender system according to user-centered criteria [28].

Offline Evaluation. In offline evaluation, recommender systems are subject to the quality criteria *accuracy*, *robustness and stability*, *coverage*, and *diversity*. To measure the *accuracy*, the predictions of the system are compared with the actual assessments of the users and the number of predictions matching the assessments are recorded. For predictions that do not match, the extent to which the values correlate with each other and the extent to which they differ is examined. The more often the predictions meet (or at least correlate with) the actual user evaluations, the higher is the *accuracy* of the recommender system. To measure *robustness and stability*, the performance of the recommender system is compared before and after a deliberate manipulation of the evaluations of individual items ("shilling attack"). If the predictions about the evaluation after the attack do not deviate strongly, the recommender system is considered *robust*. *Coverage* is high if the recommender system can access all or a very large part of the item pool for the recommendations. By contrast, systems with low *coverage* are often faced with the cold-start problem for new items. Lastly, *diversity* denotes the variability between the recommended items, i.e., how strongly the recommendations differ from each other.

To perform an offline evaluation of recommender systems, a certain number of test user ratings are removed from an existing data set. The remaining data sets are used to train the recommender system. Based on this training data, the recommender system creates recommendations for the test users or predicts ratings for the items whose ratings have been removed. The recommendations generated this way are then compared with the actual recommendations and evaluated according to the criteria presented above [28]. Offline evaluations allow for the objective evaluation of recommender systems and their underlying algorithms according to statistical criteria. Compared to online evaluations, offline evaluations are simpler and more cost-effective and they require less resources. However, as offline evaluations do not measure user satisfaction with a recommender system, some studies also include online evaluations [10,14,28]. In most cases, the recommender systems that scored best in the offline evaluation are tested by users in an online evaluation in a second step. This enables to collect the users' opinions and still save resources [10,14,28].

Online Evaluation. Online evaluations aim at obtaining a sophisticated picture of the users' attitudes. Typical concepts that are often surveyed in online evaluations include *perceived accuracy*, *perceived diversity*, *novelty*, *serendipity*, *satisfaction*, *trust*, and *data privacy concerns*. *Perceived accuracy* refers to the degree to which the users feel that the recommendations match their interests. It measures the overall assessment of the perceived quality of the recommendations [33]. *Perceived diversity* denotes how much variety the users perceive in their recommendations. When users receive the same or very similar recommendations, they may be disappointed by the recommendations and their confidence in the recommender system might decrease [33]. Most users expect a recommender system to suggest items that match their interests and preferences. If users receive item recommendations that they consider new and unexpected, and this recommendation turns out to be relevant to them, they might be positively surprised. This is called *serendipity* [28]. Closely related to this is the factor *novelty* which is about whether users perceive the recommendations of a system as new [28]. User *satisfaction* is another important dimension in evaluating recommender systems. It refers to the users' thoughts and feelings during the use of the recommender system [10,33]. User *trust* in recommender systems is influenced by both the accuracy and the transparency of the recommender system. *Trust* determines whether users rely on the recommendations or not [27,44]. Lastly, *data privacy concerns* regarding the handling of user data by recommender systems influence the willingness to release data to a recommender system. The type of data required by the recommender system and the transparency of the system influence *data privacy concerns* [27,34].

2.4 Our Research Question

As we have seen, different metrics are used in either online or offline evaluation of recommender systems. A possible downside in online evaluation is that users are asked about their perceptions after having used the recommender system. As previously mentioned this may introduce memory biases. In our approach we ask participants about their evaluation of each item directly and utilize this feedback to retrain the recommender system online. Using this approach we ask the following research question.

RQ: How do users expectations and evaluations of recommender systems depend on the accuracy of recommendations, when the recommender is trained on live feedback?

3 Method

To ensure that the participants are able to test different recommender systems "live" during the online survey, we designed a framework to establish a connection between the recommender systems and the online survey. Through this connection the participants' evaluations are forwarded directly from the survey

software to the recommender system and the items to be evaluated are dynamically provided by the recommender system. To offer suitable recommendations to the participants during the online survey, the recommender system creates a profile of each user. In this study, we tested the framework designed for evaluating recommender systems using an online survey. Following, we briefly describe the framework and the data preparation and finally describe the survey.

Framework. We designed the framework as a client-server architecture. We stored all the news articles to be recommended in a database which was connected to the server. The server hosted the recommender systems which were developed using the R package recommenderlab [19]. The client received the items from the server and showed it to the users via the Qualtrics survey that was connected to the client.

Data Preparation. As the data basis, we used the Million Post Corpus provided by Schabus, Skowron, and Trapp [37]. The database contains 12,087 articles with over one million comments published on the website of the Austrian daily newspaper *"Der Standard*[1]" from 1st of June 2015 to 31st of May 2016. To make the data usable for the recommender systems, we prepared the articles and the comments. We removed articles that were not news and therefore not suitable for the study (i.e., advertisements). We reduced the total number of articles from 12,087 to 10,309. For good readability, we limited the length of the articles to a maximum of two sections.

To ensure that no cold start problem occurs (see Sect. 2.2), artificial evaluations used to train the recommender systems were generated from the comments on the articles that contained user IDs and textual data. The comment text on the articles were converted into ratings by conducting a Sentiment Analysis using the *German sentiment vocabulary SentiWS V2.0* by Remus, Quasthoff, and Heyer [35]. If the sum of word-by-word sentiment was negative we rated the item as 1 otherwise as 5. We assigned this value as a *rating* for the article and saved it with the corresponding *Article ID* and *User ID* in a rating matrix, which we used to generate recommendations.

Online Survey. The survey consists of three parts: We first asked participants about demographic factors and attitudes. Secondly, the recommender systems were re-trained. In the third part, the users tested and evaluated two recommender systems (*UBCF* and *SVD*).

We measured all items on a six-point-Likert scale (1 - disagree very much, 2 - disagree, 3 - rather disagree, 4 - rather agree, 5 - agree, 6 - agree very much).

Demography and Attitudes. As *demographic data*, we measured *gender*, *age* and *education level* of the participants. Additionally, we measured the *computer self-efficacy (CSE)* using 8 items by Beier [5]. Moreover, we asked participants what

[1] https://www.derstandard.at/.

type of news they are interested in (*politics*, *sports*, *economics*, *culture*, *lifestyle*, *computer and technology*, *network politics* or *science*).

Expected Accuracy of Recommender Systems. We further asked the participants how accurate they expect the recommendations of certain recommender systems would be (*random*, *non-personalized*, *user-based collaborative*, *article-based collaborative*, *hybrid* and *content based*). In addition to the name of the recommender system, we described to the participants what type of recommendation the system returns and what data it requires.

Knowledge About Daily Events in Austria. Before the recommender systems are re-trained, we informed the participants that the articles shown were from the Austrian daily newspaper "*Der Standard*". We asked the participants on a six-point Likert scale (*not at all*, *not*, *somewhat not*, *somewhat*, *very* and *extremely*), how familiar they are with the daily events in Austria, as this might be a confounding variable in recommendation accuracy.

Re-training of Recommender Systems. Re-training the recommender systems is necessary to overcome the cold start problems of the UBCF and the SVD system for the participants. To re-train the recommender systems we asked participants to evaluate seven items. To select these items we need to identify informative items for evaluation. To achieve this, two other recommender systems (*RANDOM and POPULAR*) that do not have the cold start issue were used to select *five out of the seven articles*. We selected three articles *randomly* and the two *most popular*. Two further articles were selected from a set of pre-selected articles. These were chosen by three researchers by manual coding all articles based on the topics that participants indicated as favorite news topics in the survey.

Rating of Training Articles. After the participants rated these initial articles, we asked for the participant's *perception of quality*, *diversity*, *novelty*, *serendipity* and *relevance* of the recommendations, using seven statements (see Table 1). We

Table 1. Scale used for the evaluation of articles

Scale item (*perceived article quality*)	Construct
Recommended items were well chosen [a]	Quality
Recommendations differ significantly from each other [b]	Diversity
Recommendations provided new information [b]	Novelty
Recommendations surprised me [b]	Serendipity
I liked the items recommended [a]	Quality
Recommendations were relevant [b]	Relevance
Items recommended matched my interest [c]	Quality

[a]Source: Knijnenburg et al. [27], [b]Source: Fazeli et al. [16], [c]Source: Pu, Chen, and Hu [34]

used a subset of validated scales, as these questions would have to be evaluated often by the users. The new scale *perceived article quality* still shows a good internal reliability of Cronbach's $\alpha = .87$. However, we also intend to analyze on individual item levels.

Evaluation of Two Recommender Systems. Lastly, we asked the participants to also evaluate the overall performance of the *UBCF* and *SVD* recommender system. The two systems were tested one after the other. To reduce a possible sequence effect, the order in which the systems were tested was randomized. Every participant evaluated four items. Then, we asked the participants to rate the performance of the recommender systems using the scale as before (see Table 1). The scale showed a good internal reliability of Cronbach's $\alpha = .92$ for both recommender systems. We used five further statements by Knijnenburg et al. [27] (see Table 2) to ask the participants how they *assess the recommender systems*. The scale showed a good internal reliability for both systems (Cronbach's $\alpha = .88$ UBCF, $\alpha = .89$ SVD).

Table 2. List of statements used for evaluation of recommender systems

Scale item (*assessment recommender system*)
The system is useless
I would recommend the system to others
I liked the items recommended by the system
The system recommended too many bad items
I can find better items without the help of the system

Collection of Data. Participants were acquired between October and November 2019 using snowball-sampling by sending the survey via WhatsApp, Signal, Slack, and posting it on Facebook groups. We note that this yields a high social media usage bias, which we integrate when analyzing our findings.

Statistical Methods. We checked the internal reliability of the scales using the *R-package psych* [36] by calculating the Cronbach's α. We used parametric tests to check whether the results are significant. In addition, an α error of 5% ($\alpha = .05$) and a β error of 20% ($\beta = .2$) is permitted. With a sample size of $N = 140$, this means that correlations could be detected with an effect strength of $|\varrho| \geq .21$ [15].

4 Results

All procedures and statistical evaluations are available in our supplementary material in an OSF repository[2] We used R Version 3.4.1. and RMarkdown to analyze the data. After a presentation of our sample, we report our findings.

[2] https://osf.io/qn4as/.

4.1 Description of the Sample

The online survey was completed by $N = 140$ German Internet users. 64% of the users are female and the average age is 41 ($SD = 16.21$). The age distribution of the sample is bimodal. Users between 29 and 47 years are underrepresented. 41% of the users have a university degree. The users' *knowledge about the daily events in Austria* is limited ($M = 1.60$; $SD = .84$).

4.2 User Ratings of the Recommendations

Expected Quality of the Recommender Systems. Participants believe that recommender systems that *randomly* select an article recommend a suitable article with an mean accuracy of 23% ($SD = 19.24$; see Fig. 2, mean and standard error shown in red). *Popular recommender systems* are presumed to give good recommendations with an accuracy of 36% ($SD = 23.50$). An accuracy of slightly more than 50% is expected from *collaborative recommender systems*. Participants rated the *IBCF recommender system* slightly better with an accuracy of 52% ($SD = 20{,}854$) than *UBCF* with 51% ($SD = 19.34$). They rated *hybrid recommender systems* best. These should deliver good recommendations with an accuracy of 60% ($SD = 22.46$).

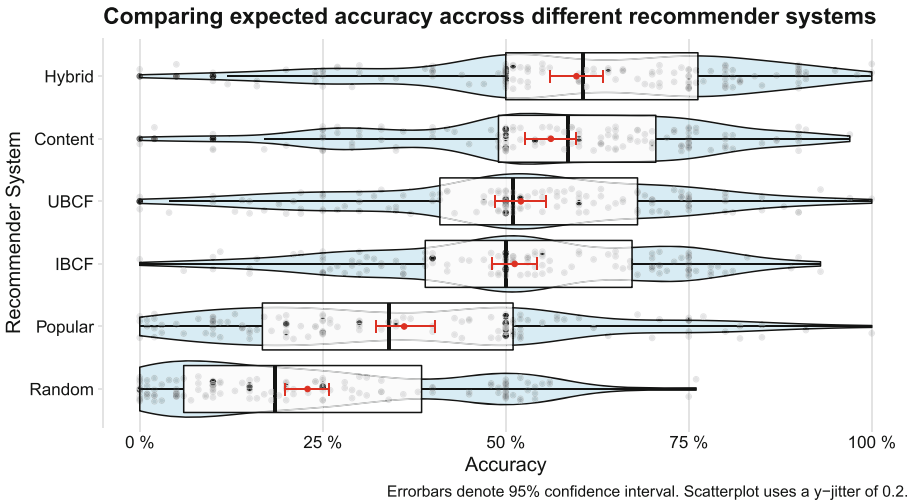

Fig. 2. How do users expect different Recommender Systems to perform?

Evaluation of the Pre-selected Articles. During the *re-training phase* (see Fig. 3), the users rated the recommendations slightly negative ($M = 3.15$; $SD = .89$). After the *re-training phase*, the users described the recommended articles as *very diverse* ($M = 4.62$; $SD = .78$). The recommended articles offered users

new information ($M = 3.79$; $SD = 1.08$), but they were not always *relevant* to them ($M = 3.45$; $SD = 1.23$). Furthermore, the users were *surprised* by the recommendations ($M = 3.69$; $SD = 1.26$). However, when users were asked again about their ratings in retrospect, users rated the *perceived quality* of the recommended articles slightly negative ($M = 3.32$; $SD = .95$).

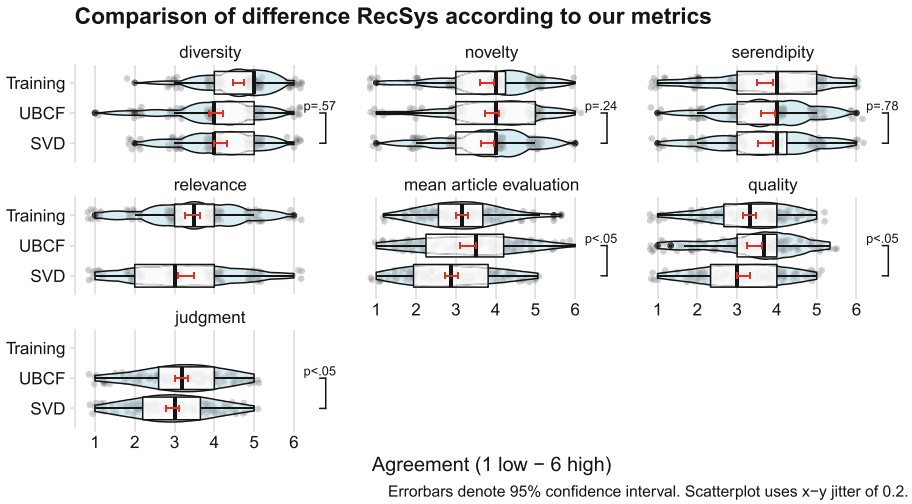

Fig. 3. Results of evaluation: comparison of different recommender systems according to our metrics

Evaluating the UBCF Recommender System. Participants rated the recommendations given by the *user-based collaborative filtering recommender system (UBCF)* slightly negative ($M = 3.31$; $SD = 1.17$; see Fig. 3). In the following overall evaluation, the users perceived the articles as *very diverse* ($M = 4.05$; $SD = 1.06$). For many users, *new information* was offered by the recommended articles ($M = 3.91$; $SD = 1.04$). The users were *surprised* by the recommendations ($M = 3.76$; $SD = 1.06$) and rated the *perceived quality* average ($M = 3.43$; $SD = 1.15$). The *overall rating* for the UBCF recommender system was slightly negative ($M = 3.17$; $SD = 1.01$) (see Fig. 3).

Evaluating the SVD Recommender System. Most of the articles recommended by the recommender system, which uses *singular value decomposition (SVD)*, were rated negative ($M = 2.87$; $SD = 1.05$). In the following *overall evaluation*, the users perceived the articles as *very diverse* ($M = 4.15$; $SD = .96$). The recommended articles provided the users with *new information* ($M = 3.79$; $SD = 1.01$), which were only of *moderate relevance* ($M = 3.28$; $SD = 1.25$). However, the users were *surprised* by the recommended articles ($M = 3.71$; $SD = 1.16$). All in all, the *perceived quality* of the recommended items ($M =3.15$;

$SD = 1.01$) was better than the individual ratings. Although, the *overall rating* of the recommender system was negative ($M = 2.94$; $SD = .96$) (see Fig. 3).

4.3 Progress of the Recommender Systems

After finishing the survey, we evaluated the improvement of the *two recommender systems*. For this, we compared the *actual article ratings* with the *predictions of the recommender systems*. We calculated the *predictions* twice. First, with the state of the recommender systems at the beginning of the survey (*pretest/ex ante*) and second, with the recommender systems after the end of the survey (*posttest/ex post*).

We tested how many article ratings the recommender systems *correctly predicted*. In the other case, we tested the extent to which the *prediction deviated* from the *actual rating*. With a good recommender system, the proportion of correct *predictions* should be as high as possible and the *deviation (RMSE)* of the remaining predictions as low as possible.

UBCF Recommender System. The *UBCF recommender system correctly predicted* every fourth item in the *pretest state* ($M = .27$; $SD = .07$). The remaining 73% of the predictions showed both under- and overestimation of up to three levels in terms of *actual rating* In the *posttest*, the system predicted every third rating correctly ($M = .33$; $SD = .05$). The over- and underestimates were still up to three levels but the *deviation* had slightly decreased. Although the distance between *prediction* and *actual evaluation* has changed in only 87 of 560 of the evaluated articles, there is a small, significant difference between *pretest* and *posttest* of the *UBCF recommender system* ($t(139.0) = 10.54, p < .001$).

The correlation between the *prediction* and the *actual ratings* improved significantly between the *pretest* and the *posttest* of the *UBCF recommender system* ($t(123) = 8.68, p < .001$). As shown in Fig. 4(a), in the *pretest* we found almost as many negative as positive correlations ($\mu = .01$) between the *predicted* and *actual ratings*. In the *posttest* we found mainly strong positive correlations ($\mu = .6$) and only sporadically strong negative correlations.

SVD Recommender System. The *pretest SVD recommender system* correctly predicted every third article rating ($M = .35$; $SD = .03$). For the remaining 65% of recommendations, there was both over- and underestimation of up to three levels between *predictions* and *actual ratings* The *posttest SVD recommender system* correctly predicted more than half of all article ratings ($M = .57$; $SD = .03$). The *deviation* between the *prediction* and the *actual evaluation* has improved for the remaining 43%. Thus, the over- and underestimation of the ratings was reduced to at most two levels. When looking at the *absolute distances*, it is evident that the difference between the *pretest* and the *posttest recommender system* is significant ($t(139.0) = 10.15, p < .001$).

The correlation between the *prediction* and the *actual ratings* improved significantly between the *pretest* and *posttest SVD recommender system* ($t(129) = 8.45, p < .001$). Figure 4(b) shows that the *pretest SVD recommender system*

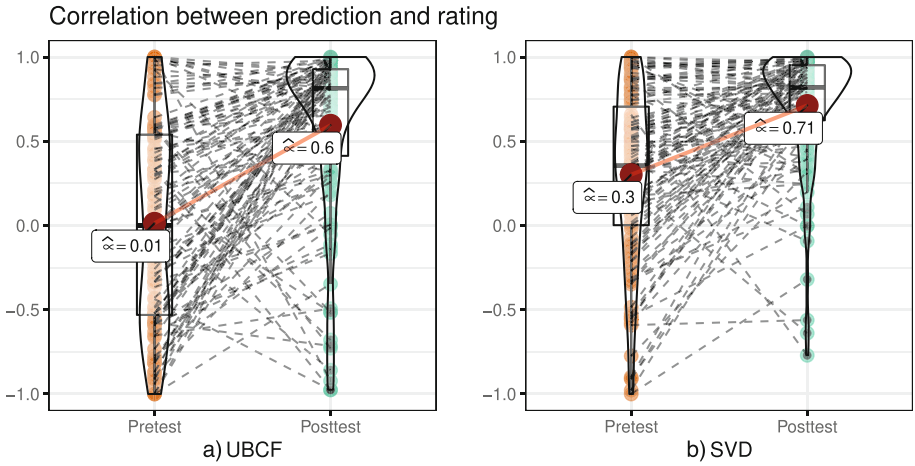

Fig. 4. Every point shown is the correlation coefficient between actual and predicted rating for a single user. Lines connect the users between pre- and post-test. Correlations above 0 indicate good predictions; correlations below 0 indicate very bad predictions. Both figures a) and b) indicate improvement, as the lines tend to move upwards. $\hat{\alpha}$ is the mean correlation coefficent for all users. The red line indicated the amount of improvement.

had more positive than negative correlations ($\mu = .3$) between the *predicted* and *actual ratings*. The *posttest SVD recommender system* shows mostly strong positive correlations ($\mu = .71$), whereas negative correlations occur only sporadically.

4.4 Relationships Between Evaluation Criteria

We analyzed how the *article ratings*, the *user-centered rating criteria*, the *expected article quality*, and the *overall rating* of the recommender system are correlated. As the functionality of a *SVD recommender system* is quite complex, we did not ask the participants which *quality* they *expected* from the *SVD recommender system*. Therefore, we evaluated the *expected quality* of the recommendations only for the *UBCF recommender system*.

Article Rating, Perceived Quality and Overall Rating. Table 3 shows that for both *recommender systems*, the *article ratings* have a significantly strong positive correlation with both the *perceived quality* and the *rating of the recommender system*. The better the users *rated the articles*, the better they *perceived the quality* of the recommendations and the better the *rating of the recommender system*. In the case of the *UBCF recommender system*, the *expected article quality* correlated slightly positively with the *actual article rating*.

User Centric Factors. For the *UBCF* and *SVD recommender system* (see Table 4) the *perceived novelty* correlates positively with the *article ratings*, the

Table 3. Correlation table of evaluation criteria

Variables	UBCF					SVD			
	M[a]	SD[b]	1	2	3	M[a]	SD[b]	1	2
1. Expected quality	51	21							
2. Article ratings UBCF	3.31	1.17	.27**			2.87	1.05		
3. Perceived quality	3.43	1.15	.16	.76**		3.15	1.01	.68**	
4. Rating	3.17	1.01	.19	.62**	.81**	2.94	.96	.57**	.80**

* indicates $p < .05$; ** indicates $p < .01$
[a] Mean, [b] Standard deviation

perceived quality and the *overall rating* of the recommender system. A user who received *novel* items through the recommended articles *rated the articles* themselves, the *quality of the recommendations* and the *entire recommender system* better than a user who did not receive novel information.

Table 4. Correlation table of user-centered criteria of the *User-based collaborative filtering (UBCF) recommender system* and *Singular value decomposition (SVD) recommender system*

Variables	UBCF					SVD					
	M[a]	SD[b]	1	2	3	M[a]	SD[b]	1	2	3	4
1. Perceived novelty	3.91	1.04				3.79	1.01				
2. Perceived diversity	4.05	1.06	.01			4.05	.96	.02			
3. Perceived serendipity	3.76	1.06	.05	.18*		3.71	1.16	.12	.24**		
4. Perceived relevance						3.28	1.25	.30**	.04	−.02	
5. Expected quality	51	21	.09	−.07	.04						
6. Article ratings	3.31	1.17	.45**	.05	−.15	2.87	1.05	.29**	.05	−.08	.41**
7. Perceived quality	3.43	1.15	.54**	.04	−.17	3.15	1.01	.45**	−.01	−.03	.54**
8. Rating	3.17	1.01	.44**	−.08	−.12	2.94	.96	.37**	−.03	−.02	.49**

* indicates $p < .05$; ** indicates $p < .01$
[a] Mean, [b] Standard deviation

Furthermore, in the *SVD recommender system*, the *perceived relevance* of the recommended articles correlates significantly and positively with the *article rating*, the *perceived quality* and the *overall rating*. The more *relevant* the recommended articles were for the users, the better the recommender systems were *rated*. This in turn positively influenced the *perceived quality* and the *overall rating of the recommender system*.

5 Discussion

In our study, we compared two recommender systems before (pretest) and after (posttest) an online study. Both recommender systems improved their prediction accuracy. Also, the correlation between the predictions and the actual ratings increased between the pre- and post-test.

Comparing the two recommender systems, the SVD recommender system scored better on all statistical measures (accuracy, deviation of predictions, and correlation between predictions and assessments) than UBCF. This suggests that users would also rate the SVD recommender system as better than the UBCF system. However, users preferred the articles suggested to them by the UBCF recommender system and perceived the quality as higher, thus giving the UBCF a better overall rating. This agrees with finding from McNee, Riedl, and Konstan [31], who found that the evaluation of a recommender system does not only depend on the accuracy of the recommendations.

We also looked at whether there is a correlation between the expected quality of the recommendations, the actual article ratings, and the subjective ratings of the users. The participants expected an accuracy of about 50% from the collaborative filtering recommender systems (UBCF). Our UBCF recommender system did not meet the expectations of the participants with an accuracy of only 33%. In contrast, the SVD recommender system exceeded expectations with 57%. The overall evaluation the recommender system shows that the perceived quality of a recommender system is correlated with the individual ratings of the recommended items.

If the participants experience the recommended items as relevant and novel, they rated the quality of the recommendations as higher and also rated these articles better. This shows that a framework like ours can help to investigate differences and relationships in algorithmic and other user-centric evaluations in online studies.

In our study, the participants had to rate each article individually, which takes a lot of time. In a next study, we want to compile the article recommendations of different recommender systems into a single generated news page after the system has been re-trained. Thus, we could compare a larger number of recommender systems. However, the simultaneous evaluation of several recommendations could also lead to problems in the comparability of the results. As news articles on news pages are often implicitly consumed, users judgments on whole pages would be influenced by many factors. Making isolation of factors harder to achieve.

In the future, the framework could be tested in other test scenarios, for example with recommender systems that recommend other products. Here, it must be considered that other types of items have different "shelf-lives" than News, therefore a drift in user preferences would have to be accounted for differently.

6 Conclusion and Outlook

In this study, we analyzed recommender systems not separately according to statistical measures, such as accuracy, or according to user-centric criteria, but to carry out both types of evaluation combined. Most interestingly, users evaluate recommendations differently than accuracy metrics, revealing the importance of studying recommender systems from a users perspective. Better accuracy does imply better user evaluation, but not solely so. Users are particularly bad

at predicting the performance of algorithmic recommendations, stressing the importance of features like explanations and visualizations of recommendations.

A balance between accuracy and user-centric criteria is nevertheless important. Our framework allows to better explore this balance and provides a starting point for further research.

Acknowledgements. This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005). ISSN 1041-4347
2. Alyari, F., Navimipour, N.J.: Recommender systems: a systematic review of the state of the art literature and suggestions for future research. Kybernetes **47**(5), 985–1017 (2018)
3. Atoum, J. O., Yakti, I.M.: A framework for real time news recommendations. In: Proceedings - International Conference on New Trends in Computing Sciences, ICTCS 2017, NJ 08854, USA, vol. 2018-Janua, pp. 89–93. Institute of Electrical and Electronics Engineers (IEEE) (2017)
4. Beck, P., et al.: A system for online news recommendations in real-time with apache mahout. In: Working Notes of the 8th International Conference of the CLEF Initiative, vol. 1866. CEUR Workshop Proceedings (2017)
5. Beier, G.: Kontrollüberzeugungen im umgang mit technik. Rep. Psychol. **9**, 684–693 (1999)
6. Bogers, T., van den Bosch, A.: Comparing and evaluating information retrieval algorithms for news recommendation. In: Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys 2007, pp. 141–144. Association for Computing Machinery (2007). ISBN 9781595937308
7. Burbach, L., et al.: User preferences in recommendation algorithms: the influence of user diversity, trust, and product category on privacy perceptions in recommender algorithms. In: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, pp. 306–310. Association for Computing Machinery (2018). ISBN 9781450359016
8. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adapt. Interact. **12**(4), 331–370 (2002). ISSN 0924-1868
9. Valdez, A.C., Ziefle, M., Verbert, K.: HCI for recommender systems: the past, the present and the future. In: Proceedings of the 10th ACM Conference on Recommender Systems, RecSys 2016, pp. 123–126. Association for Computing Machinery (2016). ISBN 9781450340359
10. Cremonesi, P., Garzotto, F., Turrin, R.: User-centric vs. system-centric evaluation of recommender systems. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013. LNCS, vol. 8119, pp. 334–351. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40477-1_21
11. De Pessemier, T., et al.: A user-centric evaluation of context-aware recommendations for a mobile news service. Multimed. Tools Appl. **75**(6), 3323–3351 (2015)

12. Duan, L., Street, N., Xu, E.: Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterp. Inf. Syst. **5**, 169–181 (2011)
13. Eirinaki, M., et al.: Recommender systems for large-scale social networks: a review of challenges and solutions. Future Gener. Comput. Syst. **78**, 413–418 (2018)
14. Ekstrand, M.D., Riedl, J.T., Konstan, J.A.: Collaborative filtering recommender systems. Found. Trends Hum.-Comput. Interact. **4**(2), 81–173 (2011). ISSN 1551-3955
15. Faul, F., et al.: Statistical power analyses using g*power 3.1: tests for correlation and regression analyses. Behav. Res. Methods **41**, 1149–60 (2009)
16. Fazeli, S., et al.: User-centric evaluation of recommender systems in social learning platforms: accuracy is just the tip of the iceberg. IEEE Trans. Learn. Technol. **PP**, 1 (2017)
17. Garcin, F., et al.: Offline and online evaluation of news recommender systems at swissinfo.ch. In: Proceedings of the 8th ACM Conference on Recommender Systems, RecSys 2014, pp. 169–176. Association for Computing Machinery (2014). ISBN 9781450326681
18. Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 257–260. Association for Computing Machinery (2010). ISBN 9781605589060
19. Hahsler, M., Vereet, B., Hahsler, M.M.: Package 'recommenderlab' (2019)
20. Hanafi, M., Suryana, N., Basari, A.S.: An understanding and approach solution for cold start problem associated with recommender system: a literature review. J. Theor. Appl. Inf. Technol. **96**(9), 2677–2695 (2018)
21. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Trans. Inf. Syst. **22**(1), 116–142 (2004). ISSN 1046-8188
22. Ishida, Y., Uchiya, T., Takumi, I.: Design and evaluation of a movie recommendation system showing a review for evoking interested. Int. J. Web Inf. Syst. **13**, 72–84 (2017). https://doi.org/10.1108/IJWIS-12-2016-0073
23. Isinkaye, F.O., Folajimi, Y., Ojokoh, B.A.: Recommendation systems: principles, methods and evaluation. Egypt. Inf. J. **16**, 261–273 (2015)
24. Karimi, M., Jannach, D., Jugovac, M.: News recommender systems - survey and roads ahead. Inf. Process. Manag. **54**(6), 1203–1227 (2018). ISSN 0306-4573. https://doi.org/10.1016/j.ipm.2018.04.008. http://www.sciencedirect.com/science/article/pii/S030645731730153X
25. Lommatzsch, A.: Real-time news recommendation using context-aware ensembles. In: de Rijke, M., et al. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 51–62. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_5
26. Kirshenbaum, E., Forman, G., Dugan, M.: A live comparison of methods for personalized article recommendation at Forbes.com. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 51–66. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33486-3_4
27. Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Model. User-Adapt. Interact. **22**(4–5), 441–504 (2012). https://doi.org/10.1007/s11257-011-9118-4. ISSN 0924-1868
28. Kotkov, D., Wang, S., Veijalainen, J.: A survey of serendipity in recommender systems. Knowl.-Based Syst. **111**, 08 (2016). https://doi.org/10.1016/j.knosys.2016.08.014

29. Li, L., et al.: Personalized news recommendation: land an experimental investigation. J. Comput. Sci. Technol. **26**, 754–766 (2011). https://doi.org/10.1007/s11390-011-0175-2
30. Lü, L., et al.: Recommender systems. Phys. Rep. **519**(1), 1–49 (2012). ISSN 0370-1573
31. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems. In CHI 2006 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2006, pp. 1097–1101. Association for Computing Machinery (2006). ISBN 1595932984. https://doi.org/10.1145/1125451.1125659
32. Phelan, O., McCarthy, K., Smyth, B.: Using Twitter to recommend real-time topical news. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys 2009, pp. 385–388. Association for Computing Machinery (2009). ISBN 9781605584355
33. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys 2011, pp. 157–164. Association for Computing Machinery (2011). ISBN 9781450306836. https://doi.org/10.1145/2043932.2043962
34. Pu, P., Chen, L., Hu, R.: Evaluating recommender systems from the user's perspective: Survey of the state of the art. User Model. User-Adapt. Interact. **22**(4–5), 317–355 (2012). https://doi.org/10.1007/s11257-011-9115-7. ISSN 0924-1868
35. Remus, R., Quasthoff, U., Heyer, G.: Sentiws - a publicly available German-language resource for sentiment analysis. In: Proceedings of the 7th International Language Resources and Evaluation (LREC 2010), pp. 1168–1171 (2010)
36. Revelle, W.R.: Psych: procedures for personality and psychological research (2017)
37. Schabus, D., Skowron, M., Trapp, M.: One million posts: a data set of German online discussions. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2017, pp. 1241–1244. Association for Computing Machinery (2017). ISBN 9781450350228. https://doi.org/10.1145/3077136.3080711
38. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative filtering recommender systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web. LNCS, vol. 4321, pp. 291–324. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72079-9_9. ISBN 9783540720782
39. Schedl, M., et al.: Current challenges and visions in music recommender systems research. Int. J. Multimed. Inf. Retrieval **7**, 95–116 (2018)
40. Shi, Y., Larson, M., Hanjalic, A.: Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges. ACM Comput. Surv. **47**(1), 1–45 (2014). https://doi.org/10.1145/2556270. ISSN 0360-0300
41. Smith, B., Linden, G.: Two decades of recommender systems at amazon.com. IEEE Internet Comput. **21**(3), 12–18 (2017). ISSN 1089-7801
42. Sohail, S.S., Siddiqui, J., Ali, R.: Classifications of recommender systems: a review. J. Eng. Sci. Technol. Rev. **10**(4), 132–153 (2017)
43. Son, L.H.: Dealing with the new user cold-start problem in recommender systems: a comparative review. Inf. Syst. **58**, 87–104 (2016). http://dblp.uni-trier.de/db/journals/is/is58.html#Son16
44. Svrcek, M., Kompan, M., Bielikova, M.: Towards understandable personalized recommendations: Hybrid explanations. Comput. Sci. Inf. Syst. **16**, 179–203 (2019). https://doi.org/10.2298/CSIS171217012S
45. Taghavi, M., et al.: New insights towards developing recommender systems. Comput. J. **61**, 319–348 (2018). https://doi.org/10.1093/comjnl/bxx056

46. Taneja, A., Arora, A.: Recommendation research trends: review, approaches and open issues. Int. J. Web Eng. Technol. **13**(2), 123–186 (2018)
47. Valdez, A.C., Ziefle, M.: The users' perspective on the privacy-utility trade-offs in health recommender systems. Int. J. Hum.-Comput. Stud. **121**, 108–121 (2019). ISSN 1071-5819. Advances in Computer-Human Interaction for Recommender Systems
48. Wang, J., de Vries, A.P., Reinders, M.J.T.: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2006, pp. 501–508. Association for Computing Machinery (2006). ISBN 1595933697
49. Zhou, X., et al.: The state-of-the-art in personalized recommender systems for social networking. Artif. Intell. Rev. **37**(2), 119–132 (2012). ISSN 0269-2821
50. Zhou, Xun, et al.: SVD-based incremental approaches for recommender systems. J. Comput. Syst. Sci. **81**(4), 717–733 (2015). https://doi.org/10.1016/j.jcss.2014.11.016
51. Özgöbek, O., Gulla, J., Erdur, C.: A survey on challenges and methods in news recommendation. In: WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies, vol. 2, pp. 278–285, January 2014