On the Use of Feature-based Collaborative Explanations: An Empirical Comparison of Explanation Styles

Sidra Naveed
University of Duisburg-Essen
Duisburg, Germany
sidra.naveed@uni-due.de

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

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

ABSTRACT

Current attempts to explain recommendations mostly exploit a single type of data, i.e. usually either ratings provided by users for items in collaborative filtering systems, or item features in content-based systems. While this might be sufficient in straightforward recommendation scenarios, the complexity of other situations could require the use of multiple datasources, for instance, depending on the product domain. Even though hybrid systems have a long and successful history in recommender research, the connections between user ratings and item features have only rarely been used for offering more informative and transparent explanations. In previous work, we presented a prototype system based on a feature-weighting mechanism that constitutes an exception, allowing to recommend both items and features based on ratings while offering advanced explanations based on content data. In this paper, we empirically evaluate this prototype in terms of user-oriented aspects and user experience against to widely accepted baselines. Two user studies show that our novel approach outperforms conventional collaborative filtering, while a pure content-based system was perceived in a similarly positive light. Overall, the results draw a promising picture, which becomes particularly apparent from a user perspective when participants were specifically asked to use the explanations: they indicated in their qualitative feedback that they understood them and highly appreciated their availability.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Collaborative Filtering; Explanations; User Experience

ACM Reference Format:
Sidra Naveed, Benedikt Loepp, and Jürgen Ziegler. 2020. On the Use of Feature-based Collaborative Explanations: An Empirical Comparison of Explanation Styles. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3386392.3399303

1 INTRODUCTION

The ever-increasing complexity of algorithms for *Recommender Systems* (RS), currently reaching its peak with the rise of deep learning, has created a demand for more explainable systems. Most research effort has been spent on explaining recommendations using only a single type of data. The most commonly used explanation style is based on *Collaborative Filtering* (CF), which only requires availability of explicit or implicit feedback provided by the user community [16]. For recommendations, ratings of those items the current user has not yet rated are predicted based on, for instance, the weighted average of ratings provided by similar users (user-based CF) or of similar items (item-based CF). Even though several attempts to explain these predictions to users have been made in the literature [10, 13, 27], this task still needs to be considered quite challenging, especially in case of model-based CF methods [14, 21, 23, 24]. A common example are the well-known Amazon "other customers also bought …" explanations, describing why an item has been recommended in a rather simpler manner—independent of the complexity of the underlying method. On the other hand, *Content-based Filtering* (CB) relies on, for example, predefined metadata or tags generated by other users or extracted from unstructured texts [9]. Accordingly, the corresponding explanation style aims at explaining the relevance of recommended items to the user's personal preferences based on content data, which thus need to be available.

Both explanation styles can be considered effective in most standard recommendation scenarios. However, recommendation processes are often more complex, especially in high risk domains such as digital cameras, cars or even houses, where purchase decisions are much more complicated than choosing a song to listen or a movie to watch. In these cases, being able to fall back on multiple datasources for providing explanations may be particularly advantageous [7]. Nonetheless, while CF has already been integrated with CB in hybrid systems, mainly to improve accuracy, current RS rarely exploit the connections between user ratings and item features for the purpose of providing richer explanations [28].

In this line of research, we in previous work proposed a hybrid approach in the domain of digital cameras that elicits the current user's preferences with respect to features of the items in a CF environment [25]. Subsequently, using a feature-weighting mechanism, similar users are computed based on these feature-based preferences instead of preferences for the items, i.e. typical ratings as in the common CF procedure (see e.g. [26]). These user-user similarities are then used to determine item recommendations based on user-item similarities calculated in the same manner. With a prototype RS built on top of this method, we showcased that we were consequently able to also use item features for providing advanced explanations of CF output in complex recommendation scenarios.

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy

In this paper, we complement this work by empirically evaluating this prototype against two established baseline systems: a standard item-based CF recommender that exclusively relies on item-item similarities in terms of ratings, and a standard CB recommender that exclusively relies on user-item similarities in terms of features, to generate both recommendations and explanations. For this, we aim at addressing the following research questions:

RQ1: Do recommendations and explanations based on feature-based CF improve subjective system assessment compared to a conventional item-based CF approach?

RQ2: Do recommendations and explanations based on feature-based CF improve subjective system assessment and user experience compared to a conventional CB approach?

2 RELATED WORK

The growing complexity of today's highly sophisticated RS algorithms has fueled the need for more transparent explanations that may help users understand the rationale behind the recommendation process [4, 17]. Accordingly, numerous approaches have been proposed for providing explanations for system-generated recommendations. However, the majority focuses on generating these explanations by exploiting only a single type of data, i.e. usually either content-related item features when CB is applied, or plain user-item ratings in a CF context [2, 8, 13, 37, 40].

Content-based explanations. RS based on CB model users by information available about the content of the items [9, 31], e.g. price, brand or color in case of digital cameras or TVs. In case of entertainment domains such as movies, genre, director and actor information may be taken into account [31]. To explain recommendations, it can consequently be made immediately clear in which sense an item is relevant, namely based on its features in comparison to the current user's feature-based profile. A prominent example is *Tagsplanations* [40], where recommended movies are explained based on tags preferred by the user. In [3], a news RS is presented that explains recommended news articles by means of keywords.

Collaborative explanations. In CF, on the other hand, user-item ratings are exploited to generate recommendations [10, 16]. Since the predictions are thus exclusively based on interaction between users and items, they can be difficult to understand. This fundamental problem of CF has already been noticed in earlier work: Herlocker et al. compared 21 different explanation styles for CF, showing that users preferred rating histograms for getting an understanding of how users with similar taste rated recommended items [13]. Later, approaches such as *PeerChooser* [27] or *SmallWorlds* [12] aimed at explaining the CF output by means of complex interactive visualizations: the active user's neighbors are displayed by connected nodes, the distance between these nodes reflects the similarity of two users. Even for model-based CF systems, visualizations have been proposed. For instance, in [11, 19], maps are used for visualizing latent item spaces, showing how user preferences are represented in the underlying model and which alternatives exist in addition to recommended items, i.e. nearby in the item space. To this day, providing explanations for CF recommendations needs however to be considered overall challenging, especially when using model-based methods: There exist of course approaches that integrate additional data, e.g. user-generated tags or aspects derived from product reviews, directly into the models for increasing transparency of recommendations and improving their fundamental explainability [14, 21, 23, 24]. Yet, the results coming from the underlying models are still abstract to a certain extent, making it difficult to present users with textual explanations that are actually meaningful. As a consequence, current real-world systems often try to explain why an item has been recommended in rather simple ways—independent of the possibly complex rationale behind these models. An example are the well-known Amazon explanations (see above) that rely entirely on plain user-item rating data. In general, such textual explanatory components are the only means that can be found in the wild for increasing the transparency of RS and their trustworthiness [38]. However, especially in light of recent findings showing that system-generated explanations are still mostly of inferior quality when compared to explanations made by humans [18], there is clearly a need for richer argumentative explanations.

Hybrid approaches. Hybrid RS have shown to benefit from both CB and CF when generating recommendations [6]. Various approaches have been proposed that use multiple datasources, some combining content and rating data [35, 36], others additionally considering social data [4, 30, 39]. However, these attempts only rarely address the goal of making the recommendation process more comprehensible. If they do, they most often make use of visualizations: An example that uses cluster maps is *TalkExplorer* [39], which allows users to explore and find relevant conference talks by analyzing connections of talks to user bookmarks, tags, and social data. *SetFusion* [30], a system based on *TalkExplorer*, instead uses Venn diagrams, yielding improvements with respect to user experience. *TasteWeights* [4] exploits social, content, and expert data to provide interactive music recommendations. The rationale behind the results is implicitly made clear by visualizing the relations between user profile, datasources and recommendations. *MyMovieMixer* [22] allows users to control the influence of different datasources in a similar fashion while immediately highlighting which filtering criteria, and to what extent, the system was able to take into account. *MoodPlay* [1] combines content- and mood-based filtering for suggesting music. The system visualizes a latent space into which both artists and moods are mapped. An avatar representing the user's profile within this visualization enables the user to comprehend why certain songs are recommended by means of their position in the latent space in relation to moods and avatar.

Summary and current work. These works have successfully introduced various interactive recommending approaches, often making results easier to understand, sometimes even via sophisticated visualizations. Yet, they usually stop short at explaining the connections between the user's preference profile and the items recommended accordingly, at most by means of his or her explicitly expressed item ratings. Content features of the items, on the other hand, could easily help provide arguments for the relevance of recommended items, allowing to provide much more informative explanations. The effectiveness of explanations that rely not only on a single type of data has been recognized [28], but received overall little attention in RS research. One of the few exceptions besides attempts that integrate additional data into model-based algorithms to foster explainability (e.g. [14, 21, 23, 24]) is our work previously

proposed in [25]. Here, CB and standard item-based CF are not only combined for improving recommendations, but for explaining the recommender's output via item features, even in complex product domains such as digital cameras. Yet, since an in-depth empirical evaluation of this explanation style from a user perspective is still missing, we aim to address this gap in the paper at hand.

3 EMPIRICAL EVALUATION

For addressing our research questions, we conducted two user studies with the goal of comparing our prototype RS for digital cameras [25] with established baselines as described in the beginning of this paper. Accordingly, we formulated the following hypotheses:

- **H1:** Feature-based CF improves subjective assessment of recommendations and explanations compared to conventional item-based CF.
- **H2:** Feature-based CF improves user experience compared to conventional item-based CF.
- **H3:** Feature-based CF improves subjective assessment of recommendations and explanations compared to conventional CB.
- **H4:** Feature-based CF improves user experience compared to conventional CB.

3.1 User study 1

The first study was conducted as a lab study with a within-subject design. To address hypotheses 1 and 2, participants were presented with two prototype systems in counter-balanced order:

- **Standard item-based CF:** Based on the item ratings each participant provided, the system recommended similar items. Explanations were provided in relation to his or her rated items, additionally showing a rating distribution graph for each of the recommended items. Participants were allowed to (re)-rate items and remove recommended items from being considered in the recommendation process (see Figure 1a).
- **Feature-based CF:** Implemented on top of our approach as described in [25], this variant closely resembled the prototype system described there as well (see [25] for more details and screenshots of the advanced explanations).

In each of the two resulting conditions, participants were first asked to indicate their preferences. In the item-based CF system, to initially provide a set of cameras to rate, we ranked available cameras by means of the *balanced strategy* based on popularity and entropy as described in [33]. The top 30 cameras were presented and participants asked to rate at least 10 of them. In the feature-based CF system, participants were asked to select at least one feature, specify values, and provide five-star ratings for all selected features.

Based on specified preferences, each system generated recommendations and corresponding explanations. Recommendations were updated immediately as soon as participants modified their preferences. They had unlimited time to explore the respective system variant. We asked them to explore each recommended camera and its corresponding explanation, compare it with their indicated preferences, and select cameras that satisfy their needs by adding them to a shopping basket. Participants could finish the task at their own discretion after selecting at least one camera. After exploring the system and making their decisions, they had to evaluate the system subjectively and fill in a questionnaire. On average, participants spent 30–35 minutes to complete the study.

Participants and questionnaire. 20 students (14 females) with age of $M = 26.45$, $SD = 3.00$ (range 21–40 years) participated in the study. They were rewarded with study credit for participation.

The questionnaire used for the subjective assessment was primarily based on the pragmatic procedure for evaluating RS proposed in [15]. We assessed *Perceived Recommendation Quality*, *Choice Satisfaction*, and *Usage Effort* by means of this framework. To assess *Explanation Quality*, *Transparency*, *Trust*, and *Overall Satisfaction*, we used items from [32]. We further wanted to explore how the novel feature-based explanations support users in their decision-making process. This aspect is usually not studied in-depth in RS user studies, but may be measured in terms of several factors such as ease of understanding, helpfulness, appropriateness, and information sufficiency. Thus, we created items ourselves similar to [34] for measuring the impact of explanations on *Decision Support*.

Since our prototype with its additional feature-based explanations also provided richer interaction possibilities, we investigated user experience in contrast to the item-based CF system. For this, we used the *System Usability Scale* (SUS) [5] and the *Interface Adequacy* construct from [32]. All questionnaire items were rated on a 1–5 Likert response scale. Additionally, for qualitative feedback, we provided open-ended questions, asking participants which system out of the two they preferred and for which reasons. Moreover, we asked participants to report suggestions or complaints regarding their preferred system and its functionality. Interaction logs were recorded in terms of number of clicks on certain areas of interest.

Hypothesis 1. First, we conducted a one-way repeated measures MANOVA ($\alpha = 0.05$), revealing statistically significant differences between item- and feature-based CF for aggregated dependent variables, $F(8, 13) = 3.39$, $p = .025$, $\eta_p^2 = .67$, Wilk's $\lambda = 0.324$. To determine individual effects regarding specific dependent variables, we ran univariate tests. The results shown in Table 1 indicate that for most variables, the feature-based CF system received higher scores than the other. Differences were significant for condition in terms of *Perceived Recommendation Quality*, *Explanation Quality*, *Transparency*, *Satisfaction*, and *Decision Support*. However, we did not find significant effects for *Choice Satisfaction*, *Usage Effort*, and *Trust*. Yet, results overall indicate that we can generally accept H1.

Table 1: Mean values and standard deviations *(df=20)* for the subjective system assessment of the different conditions. Significant differences are marked by *. Higher values (highlighted in bold) indicate better results.

Construct	Item-based CF		Feature-based CF				
	M	*SD*	*M*	*SD*	*F*	*p*	η_p^2
Perc. Rec. Quality	3.52	0.82	**4.00**	0.57	5.36	**.031***	.212
Choice Satis.	**4.09**	0.62	4.04	0.80	0.08	.771	.004
Usage Effort	3.66	0.84	**3.80**	0.53	0.40	.531	.020
Expl. Qual.	2.71	1.23	**4.47**	0.60	29.7	**<.001***	.598
Transparency	2.76	1.04	**4.23**	1.04	23.3	**<.001***	.538
Trust	3.19	0.67	**3.52**	0.74	3.68	.069	.156
Decision Support	3.07	0.57	**3.73**	0.44	12.8	**.002***	.391
Overall Satis.	3.23	0.94	**3.90**	0.70	7.0	**.016***	.259

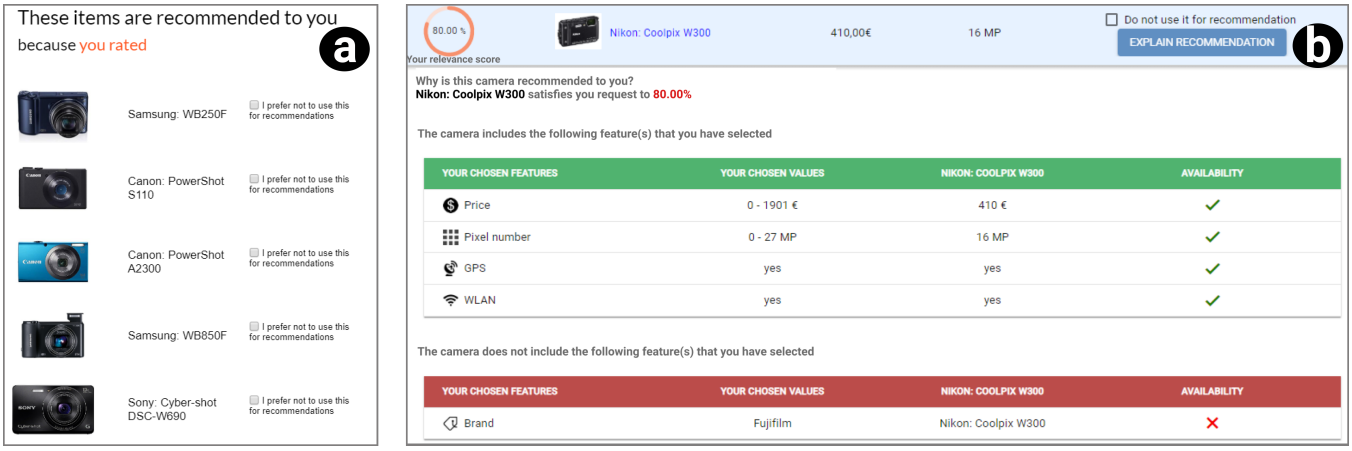

Figure 1: Screenshots of the baselines systems: a) Item-based CF recommender with explanations based on ratings for similar items, b) CB recommender with explanations of recommended items in terms of features based on user-item similarities.

Hypothesis 2. A paired t-test indicated no statistical difference between feature-based CF ($M = 4.16$, $SD = 0.54$) and item-based CF ($M = 3.91$, $SD = 0.66$) with respect to *Interface Adequacy* ($t(20) = 1.66$, $p = .11$). However, although the SUS yielded no significant difference ($t(20) = 1.55$, $p = .13$), we observed a tendency in general usability towards feature-based CF, with a SUS score of 74, as opposed to only 67 for item-based CF. Nevertheless, we cannot accept H2.

Log data for item-based CF showed that all participants explored item recommendations and their explanations. In feature-based CF, item recommendations and corresponding explanations were relatively more often explored than feature-based equivalents.

Moreover, we explicitly asked participants about their preferred system: 16 out of 20 preferred feature-based CF. In answers to the open-ended question, the majority indicated to like the option to specify preferences in terms of features, availability of additional feature-based information regarding the products, and explanations in form of direct comparisons of recommended cameras with preferred features. One participant wrote that he or she liked the "amount of information" and the "explanation why an object was recommended". Another participant said that he or she liked the system because it allows "selecting certain camera features that are important to you and evaluating how important they actually are".

3.2 User study 2

The second study designed to test hypotheses 3 and 4 was conducted via Amazon Mechanical Turk and consisted of two parts. The first part had a between-subject design in which participants were presented with one of the two following prototype systems:

- **Standard CB:** Recommendations and explanations were generated based on features and values selected by participants. The system allowed adding or removing features to manipulate recommendations. Explanations were shown for each recommended item in form of a comparison table of item features and preferred features (see Figure 1b).
- **Feature-based CF:** See user study 1, or [25], respectively.

In each of the resulting two conditions, participants were first asked to indicate their preferences in terms of features. After the system generated corresponding recommendations, they were required to explore recommendations and explanations, using the interactive mechanisms provided, in order to get an understanding of the rationale behind the results and of the relevance of recommended items in relation to their personal preferences. After finishing the task at their own discretion, participants were asked to subjectively evaluate the system and to fill in a questionnaire.

The second part was conducted in a similar fashion, but with only a single condition using our feature-based prototype. Instead of just exploring recommendations and explanations, we laid the focus of participants specifically on the additionally provided explanation facilities: We explicitly asked them to explore each kind of explanations provided by the system (*similar users*, *item recommendations*, *feature recommendations*; see [25]) in order to understand these explanations and answer corresponding questions later on.

Participants and questionnaire. For the first part, a total of 100 participants were recruited. They received a reward of 1 USD for completing the study, which took approximately 15–20 minutes. After excluding incomplete responses, outliers, and participants who did not take the experiment seriously, 73 participants (27 females) with an age of $M = 38.32$, $SD = 8.64$ (range 25–66 years) were considered for further analysis. Out of these, 35 were assigned to the CB and 38 to the feature-based CF condition. For the second part, we recruited 40 participants. 1 USD was given as an incentive for completion, which again took 15–20 minutes. After excluding outliers and incomplete responses, we used data from 32 participants (12 females) with an age of $M = 35.81$, $SD = 8.18$ (24–60 years).

For subjectively evaluating the two systems, we used basically the same questionnaire as in study 1. For user experience, we additionally took the short version of the *User Experience Questionnaire* (UEQ) [20] into account (7-point bipolar scale ranging from -3 to 3). For qualitative feedback, open-ended questions were asked regarding preferences, suggestions and complaints about the respective system and its functionality. In addition, interaction logs were recorded. In the second part of the study, we extended the questionnaire by asking participants specific questions about the explanation facilities while they performed the task. With these

questions, we wanted to specifically find out whether participants understood the explanations and how they perceived the additionally available feature-based mechanisms.

Hypothesis 3. First, we again conducted a MANOVA ($\alpha = 0.05$) to study possible differences between the two systems. The analysis revealed statistically significant differences between the CB and feature-based CF system for aggregated dependent variables, $F(8, 64) = 2.63$, $p = 0.015$, $\eta_p^2 = 0.24$, Wilk's $\lambda = 0.75$. Table 2 shows that the CB system received better results than our feature-based CF system with respect to most of the dependent variables. Yet, univariate tests indicated statistical significance only for *Decision Support* and *Overall Satisfaction*. For the other constructs, *Perceived Recommendation Quality*, *Choice Satisfaction*, *Usage Effort*, *Explanation Quality*, *Transparency*, and *Trust*, differences were negligible. Overall, the results however indicate that we cannot accept H3.

Table 2: Mean values and standard deviations *(df=71)* for the subjective system assessment of the different conditions. Significant differences are marked by *. Higher values (highlighted in bold) indicate better results.

Construct	Content-based		Feature-based CF				
	M	*SD*	*M*	*SD*	*F*	*p*	η_p^2
Perc. Rec. Quality	**4.18**	0.68	4.09	0.62	3.72	.544	.005
Choice Satis.	**4.25**	0.65	4.02	0.67	2.17	.145	.030
Usage Effort	3.45	0.63	**3.69**	0.84	1.86	.176	.026
Expl. Quality	**4.37**	0.80	4.26	0.86	0.30	.582	.004
Transparency	**4.42**	0.69	4.23	0.78	1.20	.276	.017
Trust	**4.08**	0.98	4.02	0.88	0.07	.786	.001
Decision Support	**4.01**	0.50	3.71	0.66	4.70	**.034***	.062
Overall Satis.	**4.40**	0.65	4.00	0.98	4.10	**.046***	.055

In the second part of the study, we evaluated the feature-based system once again. Now with only a single condition, we laid participants' focus on the additional feature-based interaction possibilities and explanation facilities. This was done by changing the task description and adding corresponding open-ended questions so that participants were really required to use these mechanisms more extensively. Table 3 shows the results in comparison to the assessment of the feature-based CF approach in the first part (i.e. as also shown in the previous table). We found no significant differences in the assessment of the feature-based CF system without (part 1) and with (part 2) a specific task focusing on explanations.

Table 3: Mean values and standard deviations for the subjective system assessment of feature-based CF without (part 1) and with (part 2) a task focusing on explanations.

Construct	Part 1		Part 2	
	M	*SD*	*M*	*SD*
Perc. Rec. Quality	4.09	0.62	4.00	1.01
Choice Satisfaction	4.02	0.67	4.09	0.92
Usage Effort	3.69	0.84	3.54	0.99
Explanation	4.26	0.86	4.28	0.88
Transparency	4.23	0.78	4.46	0.91
Trust	4.02	0.88	3.93	1.01
Decision Support	3.71	0.66	3.79	0.83
Overall Satisfaction	4.00	0.98	4.25	1.13

Hypothesis 4. An independent *t*-test showed no statistical significant difference with respect to *Interface Adequacy* ($t(71) = .80$, $p = .33$) between CB ($M = 4.12$, $SD = 0.52$) and feature-based CF ($M = 4.00$, $SD = 0.80$). More in-depth usability evaluation also did not reveal any significant effects: SUS score was 75 in the CB, but still 71 in the feature-based CF condition ($t(71) = .73$, $p = .46$). With respect to the different subscales of the UEQ, the CB system received scores of: 2.15 for pragmatic quality (excellent), 1.30 for hedonic quality (above average), and 1.73 overall (good). The feature-based CF system received slightly lower scores: 1.87 for pragmatic quality (excellent), 1.06 for hedonic quality (below average), and 1.47 overall (good). We found similar results when evaluating only the feature-based CF system in the second part: SUS score was 66, not significantly different from the first part ($t(68) = .98$, $p = .33$). The scores
quality (above average), 1.08 for hedonic quality (above average), and 1.2 overall (above average). Thus, we can overall not accept H4.

Log data recorded for both systems did not provide results worth reporting. In case of CB, when asked about the functionality they liked the most, the majority of participants responded with: the option of adding more features from a drop-down list, and how the system provided different features to select from. They also liked the explanations: one participant indicated that "the explanation was helpful as it provided somewhat detailed information about the camera's features and why it was recommended". Another participant enjoyed "the explanation, as it quickly showed the features one was looking for". In case of the feature-based CF system (in part 1 of the study), the majority liked the option of rating features they selected to indicate how important they considered them. Participants appreciated, for example, "the rating of different features since it allowed prioritizing the features [they] really liked" and that they were "able to show how important particular features were, and not just that [they] wanted to consider them".

4 DISCUSSION

In the first user study, we investigated the impact of our feature-based CF approach on subjective system assessment and user experience in comparison to a conventional item-based CF approach (hypotheses 1 and 2). The results show that the prototype system that relied on feature-based CF received significantly higher scores compared to the item-based CF variant in a number of dependent variables. Even in cases where differences were not significant, feature-based CF tended to yield better results. Only regarding satisfaction with their choice, participants provided slightly higher ratings in the item-based condition. While this indicates that they were in the end able with both systems to successfully choose an item that satisfies their needs, the superiority in all other aspects emphasizes that the process of getting there is perceived much better when the system additionally takes item features into account. This is also reflected by the significantly higher overall satisfaction.

Additionally, we compared the two systems with respect to aspects related to user experience, in particular, interface adequacy and general usability. Although analyses yielded no statistical significant differences, it seems noteworthy that feature-based CF always performed better than the baseline. This indicates that regardless of the more advanced interaction mechanisms and the richer explanations provided in the feature-based CF system compared to the item-based CF system (where only much simpler explanations based on item ratings were presented, without considering features

and preferences for these features), participants perceived both systems to be of similar quality in terms of usability.

We also asked participants about their preferred system and the reasons for preferring it. The majority indicated that they preferred the system which relied on our feature-based CF approach, and most of them explicitly wrote that they liked the explanations as well as the possibility to directly compare features of recommended cameras with their preferred features. For instance, one participant stated that he or she "liked the comparison of the properties of the recommended camera with the selected properties". Most participants also appreciated the option to provide five-star ratings for explicitly indicating the individual importance of selected features.

In the second user study, we aimed at evaluating our feature-based CF approach in comparison to a standard CB recommender (hypotheses 3 and 4). In the first part, where we directly compared the two approaches, there were tendencies in favor of CB, but we found only few significant effects, i.e. the systems were in general assessed similarly positively. Only in terms of decision support and overall satisfaction, CB received significantly higher ratings. Again without significance, the CB system tended to perform better with respect to UEQ results as well. On the other hand, qualitative feedback provided in the feature-based CF condition emphasizes similar to study 1 that participants appreciated the option to rate features—which is not even possible with conventional CB. Additionally, they liked the explanations in form of one-to-one comparisons of features of recommended items with their preferred features. Other explanation facilities (i.e. explanations of feature recommendations and based on similar users) apparently received less attention. Still, one participant wrote that he or she liked "the option of having explanations based on experience of other users". A reason that the other options were not used to a larger extent might be the limited task description in the first part: Participants were able to finish system usage at their own discretion as soon as they found at least one camera in accordance with their indicated preferences. Thus, considering that we recruited our sample from Amazon MTurk, and that there was no need to make an actual purchase decision, they might have seen no need to interact more thoroughly with explanations other than the one-to-one comparisons. On the contrary, it was possible to complete the task without putting additional effort in exploring the advanced explanation facilities and spending more time for understanding them. We assume this would be different in a real-world setting, when facing an actual purchase situation.

In line with these findings, results from the second part of the study, which put emphasis on the additional feature-based options, particularly highlight that participants not only understood, but also appreciated the novel explanation facilities (i.e. for feature recommendations and based on similar users). One participant suggested that "explanations based on similar users would probably be helpful to someone who wants to see how others who have similar preferences would go about selecting a camera". Another participant indicated that he or she "wants to see the link to specific users to further explore their profile or to see more about them".

Nevertheless, we want to remark that the first part of study 2 was in contrast to study 1 conducted in a between-subject design to avoid carry-over effects. Thus, the vastly positive reception of the system in the CB condition might be explained under the assumption that for this more simple variant, participants were still provided with everything needed and probably also expected from such kind of RS: recommendations in this case are simply the result of a content-based recommendation procedure (i.e. solely relying on item features), making additional interaction mechanisms and especially more complex explanations not only hardly possible to implement, but also less meaningful. The gain through the integration of item features becomes more apparent from a user perspective when comparing against approaches that usually do not take such content-related data into account: Concretely, when compared to standard item-based CF as in the first study, the value is in contrast to a CB approach that inherently uses these features for generating recommendations immediately clear. At the same time, CB approaches cannot benefit from the advantages provided by CF, including the consideration of rating data from other users, which is useful especially in the long run (but therefore quite difficult to investigate in typical user studies). Instead, using CB only, the risk increases of soon getting stuck in a filter bubble [29].

Overall, the results from the second study thus also contribute to the positive image of our extended CF approach: Apparently, the increased complexity does not considerably affect participants impression in a negative way. In fact, in all comparisons, results are still very promising. Nevertheless, it is worth mentioning that the feature-based CF approach needs improvement. For example, some participants indicated that "the system needs to be simplified" or that "the interface should make comparisons easier by placing the different options side by side instead of having to click on things individually". Yet, such comments (which we received a few in both user studies) are more related to general usability issues, but addressing them is still an important aspect of our future work.

5 CONCLUSION & OUTLOOK

In this paper, we investigated the use of advanced feature-based CF style explanations in the complex domain of digital cameras, based on our method proposed in [25] for integrating user ratings and item features in a hybrid fashion. To study the impact from a user perspective, we compared our prototype RS with established baselines, a conventional item-based CF and a CB system. The results of two empirical studies show that feature-based CF performs (significantly) better than item-based CF in terms of almost all aspects (RQ1). Participants appreciated that additional features were taken into account, which is consistent with earlier work (e.g. [16, 21]). On the other hand, when comparing our approach with CB, most differences were in favor of CB, if only significant in terms of two aspects. This seems to indicate that the higher complexity of the feature-based CF approach does not negatively affect user perception. In contrast, the absence of considerable differences in terms of user experience indicates that participants perceived the system to be of at least similar quality when compared to much simpler but established baselines—which especially in tandem with the qualitative feedback draws a very positive picture (RQ2). However, we believe that amongst other factors, the complexity of domain and decision task might affect user's need to see explanations at certain levels of detail, hence possibly interfering with the perception of the system, and thus requiring further investigation in future work.

REFERENCES

[1] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2016. Moodplay: Interactive mood-based music discovery and recommendation. In *UMAP '16: Proceedings of the 24th ACM Conference on User Modeling Adaptation and Personalization*. ACM, New York, NY, USA, 275–279.

[2] Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop*, Vol. 5. 153.

[3] Daniel Billsus and Michael J. Pazzani. 1999. A personal news agent that talks, learns and explains. In *AGENTS '99: Proceedings of the 3rd Annual Conference on Autonomous Agents*. ACM, New York, NY, USA, 268–275.

[4] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A visual interactive hybrid recommender system. In *RecSys '12: Proceedings of the 6th ACM Conference on Recommender systems*. ACM, New York, NY, USA, 35–42.

[5] John Brooke et al. 1996. SUS—A quick and dirty usability scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[6] Robin Burke. 2007. Hybrid Web Recommender Systems. In *The Adaptive Web. Methods and Strategies of Web Personalization*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Lecture Notes in Computer Science, Vol. 4321. Springer, Berlin, Germany, 377–408.

[7] Jorge Castro, Rosa M. Rodriguez, and Manuel J. Barranco. 2014. Weighting of features in content-based filtering with entropy and dependence measures. *International journal of computational intelligence systems* 7, 1 (2014), 80–89.

[8] Shuo Chang, F. Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 175–182.

[9] Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2015. *Recommender Systems Handbook*. Springer US, Boston, MA, USA, Chapter Semantics-aware content-based recommender systems, 119–159.

[10] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. 2011. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2011), 81–173.

[11] Emden R. Gansner, Yifan Hu, Stephen Kobourov, and Chris Volinsky. 2009. Putting recommendations on the map: Visualizing clusters and relations. In *RecSys '09: Proceedings of the 3rd ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 345–348.

[12] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. 2010. SmallWorlds: Visualizing social recommendations. *Computer Graphics Forum* 29, 3 (2010), 833–842.

[13] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. [n.d.]. Explaining collaborative filtering recommendations. In *CSCW '00: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*.

[14] Yunfeng Hou, Ning Yang, Yi Wu, and S Yu Philip. 2019. Explainable recommendation with fusion of aspect information. *World Wide Web* 22, 1 (2019), 221–240.

[15] Bart P. Knijnenburg, Martijn C. Willemsen, and Alfred Kobsa. 2011. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender systems*. ACM, New York, NY, USA, 321–324.

[16] Yehuda Koren and Robert M. Bell. 2015. *Recommender Systems Handbook*. Springer US, Boston, MA, USA, Chapter Advances in collaborative filtering, 77–118.

[17] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 379–390.

[18] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and JÃijrgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *CHI '19: Proceedings of the 37th ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.

[19] Johannes Kunkel, Benedikt Loepp, and Jürgen Ziegler. 2017. A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering. In *IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 3–15.

[20] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, 63–76.

[21] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive recommending with tag-enhanced matrix factorization (TagMF). *International Journal of Human-Computer Studies* 121 (2019), 21–41.

[22] Benedikt Loepp, Katja Herrmanny, and Jürgen Ziegler. 2015. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *CHI '15: Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 975–984.

[23] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Coevolutionary Recommendation Model: Mutual Learning between Ratings and Reviews. In *WWW '18: Proceedings of the 2018 World Wide Web Conference*. International WWW Steering Committee, Geneva, Switzerland, 773âĂŞ782.

[24] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *RecSys '13: Proceedings of the 7th ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 165–172.

[25] Sidra Naveed and Jürgen Ziegler. 2019. Feature-driven interactive recommendations and explanations with collaborative filtering approach. In *ComplexRec '19: Proceedings of the 3rd Workshop on Recommendation in Complex Scenarios*.

[26] Xia Ning, Christian Desrosiers, and George Karypis. 2015. *Recommender Systems Handbook*. Springer US, Boston, MA, USA, Chapter A comprehensive survey of neighborhood-based recommendation methods, 37–76.

[27] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: Visual interactive recommendation. In *CHI '08: Proceedings of the 26th ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1085–1088.

[28] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24, 3 (2012), 555–583.

[29] Eli Pariser. 2011. *The filter bubble: What the internet is hiding from you*. Penguin Press, New York, NY, USA.

[30] Denis Parra, Peter Brusilovsky, and Christoph Trattner. 2014. See what you want to see: Visual user-driven approach for hybrid recommendation. In *IUI '14: Proceedings of the 19th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 235–240.

[31] Michael J. Pazzani and Daniel Billsus. 2007. Content-Based Recommendation Systems. In *The Adaptive Web. Methods and Strategies of Web Personalization*, Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl (Eds.). Lecture Notes in Computer Science, Vol. 4321. Springer, Berlin, Germany, 325–341.

[32] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender systems*. ACM, New York, NY, USA, 157–164.

[33] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. 2002. Getting to know you: Learning new user preferences in recommender systems. In *IUI '02: Proceedings of the 7th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 127–134.

[34] María Luisa Sanz de Acedo Lizarraga, María Teresa Sanz de Acedo Baquedano, Maria Soria Oliver, and Antonio Closas. 2009. Development and validation of a decision-making questionnaire. *British Journal of Guidance & Counselling* 37, 3 (2009), 357–373.

[35] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2008. Justified recommendations based on content and rating data. In *WebKDD Workshop on Web Mining and Web Usage Analysis*.

[36] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2008. Providing justifications in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, 6 (2008), 1262–1272.

[37] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2009. MoviExplain: A recommender system with explanations. In *RecSys '09: Proceedings of the 3rd ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 317–320.

[38] Nava Tintarev and Judith Masthoff. 2015. *Recommender Systems Handbook*. Springer US, Boston, MA, USA, Chapter Explaining recommendations: Design and evaluation, 353–382.

[39] Katrien Verbert, Denis Parra, and Peter Brusilovsky. 2014. The effect of different set-based visualizations on user exploration of recommendations. In *IntRS '14: Proceedings of the 1st Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 37–44.

[40] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining recommendations using tags. In *IUI '09: Proceedings of the 14th International Conference on Intelligent user interfaces*. ACM, New York, NY, USA, 47–56.