Exploring Multi-List User Interfaces for Similar-Item Recommendations

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

Mathias Jesse
University of Klagenfurt, Austria
mathias.jesse@aau.at

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

Christoph Trattner
University of Bergen, Norway
christoph.trattner@uib.no

ABSTRACT

On many e-commerce and media streaming sites, the user interface (UI) consists of multiple lists of item suggestions. The items in each list are usually chosen based on pre-defined strategies and, e.g., show movies of the same genre or category. Such interfaces are common in practice, but there is almost no academic research regarding the optimal design and arrangement of such multi-list UIs for recommenders. In this paper, we report the results of an exploratory user study that examined the effects of various design alternatives on the decision-making behavior of users in the context of similar-item recommendations. Our investigations showed, among other aspects, that decision-making is slower and more demanding with multi-list interfaces, but that users also explore more options before making a decision. Regarding the selection of the algorithm to retrieve similar items, our study furthermore reveals the importance of considering social-based similarity measures.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Recommender System, User Interface, User Study

ACM Reference Format:
Dietmar Jannach, Mathias Jesse, Michael Jugovac, and Christoph Trattner. 2021. Exploring Multi-List User Interfaces for Similar-Item Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3450613.3456809

1 INTRODUCTION

Many modern online services, such as e-commerce or media streaming sites, provide automated recommendations to their users. In many cases, these recommendations are presented to users in multiple lists in the user interface (UI). The web pages of services like Amazon or Netflix, for example, are dominated by such lists that contain item suggestions. In such a UI, each list shows items that are presented according to pre-defined criteria, e.g., based on the genre or category or because they are a good match with the customer's recent preferences. Such multi-list interfaces can serve different purposes. They may, for example, help users discover new items or categories they were unaware of. At the same time, they may facilitate a more efficient exploration process as users are able to scan a relevant subset of the available options more quickly.

While such UIs are very common in practice, almost no academic research regarding their design exists, even though it is well known that not only the selection of items determines the effectiveness of a recommender system, but also the design of the user experience [6, 9, 12]. A typical open question in the context of multi-list designs is how the lists should be selected and ordered. One could also ask if one item should appear in more than one list. Finally, it is also not entirely clear if such interfaces would lead to more exploration or a more efficient decision-making process. One of the few works that discuss some of these questions was published by Netflix [5]. Their goal is to automatically select and rank the lists to be displayed to users in a personalized way while at the same time observing a number of constraints. Unfortunately, little is revealed in the paper regarding the effects of different page arrangements on users.

With this work, our goal is to shed light on design questions for multi-list UIs for recommender systems and to contrast multi-list interfaces with traditional single-list representations. As an application example, we focus on the similar-item recommendation problem [13, 14]. Today, such recommendations are implemented through multi-list interfaces at popular services including Netflix or Amazon Prime Video, where users can browse lists of movies of the same genre or feature the same actors as a reference movie.

To answer our research questions, e.g., regarding the selection of lists or the effects on the decision-making process, we created various alternative multi-list designs, which we evaluated in a controlled study with the help of crowdworkers (N=775). Among other aspects, our exploratory study indicates that a careful selection of lists is important and that multi-list interfaces can in general lead to more explorative, yet slower decision making processes.

2 PREVIOUS WORK

We discuss previous research on the two main topics of our study, multi-list interfaces and similar-item recommendations. In both areas, the literature is scarce despite their practical relevance.

Multi-list Recommendation Interfaces. While there are various ways of how to design the details of the UI of a recommender system

UMAP '21, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8366-0/21/06…$15.00
https://doi.org/10.1145/3450613.3456809

[6], almost all academic research in the area implicitly assumes that one single item list is shown to the users. This is surprising, given the ubiquitous use of multiple lists in practice.

In [1, 5], the authors discuss the page generation system of Netflix. One main design goal is to present a diverse selection of content to accommodate, e.g., a user's changing mood from session to session and to ease catalog navigation. After experimenting with rule-based approaches, their solution was an algorithmic one that optimizes the relevance of individual lists for users while maintaining overall diversity and taking device constraints into account. To that purpose, a number of features were extracted for the supervised learning task, and suitable offline metrics were designed to assess the quality of the generated pages before A/B testing. In their works, unfortunately no insights were shared regarding what makes a page layout effective and how different layouts impact user behavior. With our work, our goal is to investigate such questions with the help of a user study.

In a recent report [8], insights regarding the automated optimization of the multi-list landing pages of a major travel platform were reported. In this application, it is particularly important to provide recommendations that match the customers' state in their shopping journey. Similar to our work, the authors of [8] tried different strategies to select the lists, e.g., based on the user's recent search activities or based on a given reference item. Their results, like ours, show that the choice of lists can make a strong difference in terms of different performance objectives.

One example for the use of multiple lists in academic research is [10]. Here, the authors explore the use of *trust-inspiring* user interfaces for critiquing-based recommenders. In their UI, item suggestions are presented in multiple lists, where each list contains items that represent trade-offs to a reference item, e.g., *"cheaper and lighter, but lower processor speed"*. This so-called *organization interface* was compared to a single-list UI in a user study, and the authors found that the multi-list interface for example led to higher perceived competence and lower perceived effort (but not to a lower task completion time). Our study is similar as we also aim to understand the impact of different UI variants on user behavior and user perceptions. Differently from [10], we however do not focus on questions of trust or assumed competence but on practical design questions.

Similar-Item Recommendation. Similar-item recommendations are common in practice, and they are sometimes also organized in multiple lists, e.g., on media streaming sites. In the literature, only a few works exist, and none of them considers multi-list UIs. In [15], the use of human similarity judgments was explored to build a content-based recommendation approach, e.g., based on movie genres. A related study was performed in [14], where the authors collected over 20,000 pairwise similarity judgments as a basis for their research. The goal of their study was then to identify which recommendation strategy is able to match these human perceptions best. Finally, instead of applying existing algorithms, the authors of [13] aimed to learn a recommendation function from the data by correlating item features (e.g., movie genres) with human judgments. Overall, while the focus of each work [13–15] was different, they in particular indicate that collaborative signals (e.g., ratings) can be good predictors of similarity besides the content features. We base some of our experiment designs on these insights.

3 METHODOLOGY

We conducted a between-subjects online study, where participants were, after informed consent, tasked to *(i)* search for a movie that they recently watched and enjoyed, *(ii)* pick one system-recommended movie as a similar one to watch next, and *(iii)* provide answers to a post-task questionnaire (Figure 1).

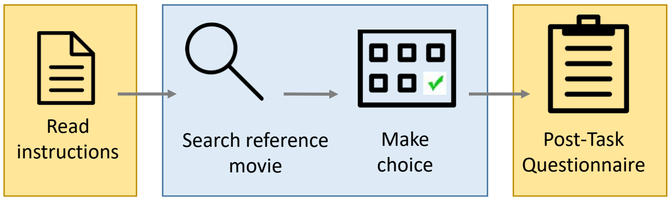

Figure 1: Experiment flow

Investigated UI Designs. We developed a web application for the purpose of the study. After choosing a reference movie, each participant was shown 30 similar-movie recommendations[1]. We created eight alternative ways of presenting similar items in terms of content and layout. We used two basic visual structures as shown in Figure 2. Screen captures of our application, of commercial solutions, and additional materials can be found online[2]. In the *single-list condition*[3], the 30 recommendations were presented as one long list with line breaks, whereas in the *multiple list* condition, each list had an own label, thus indicating that the items were selected based on different criteria. When users hovered with the mouse over an item, additional information was provided, e.g., the title or the release year. After a click on an item, even more information (e.g., about the actors or directors) was shown, and participants could then select the movie as their choice.

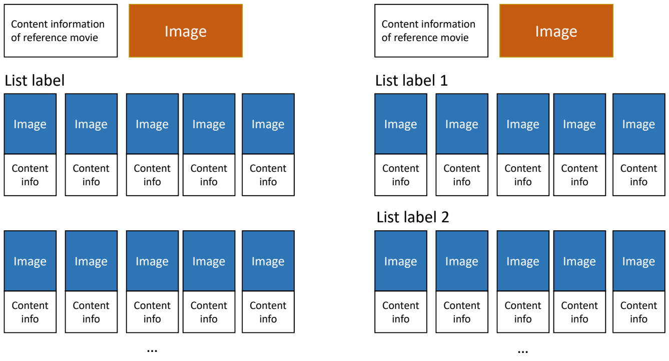

Figure 2: Single-list (left), multiple-list (right) layouts

[1] We chose 30 movies to provide a sufficiently large choice set that does not overwhelm participants.

[2] https://github.com/MathiasJesse/ACM-UMAP-2021-Submission-Multi-List-Recommendation-Online-Material

[3] While the single-list interface is also organized in multiple rows due to the line breaks, we assume it is viewed as one single ranked list as there is only one list label.

The UI designs were not only varied in terms of the layout, but also in terms of how the similar-item suggestions were determined. Details about the eight designs, i.e., which algorithms we used for selecting items and how they were presented, are shown in Table 1. We applied the algorithms on the MovieLens dataset, additional content information and human judgments shared in [13]. The rationale for the different designs will be laid out in the next section.

Dependent Variables. In the post-task questionnaire, we asked participants 17 questions (details provided in the online material), e.g., about choice difficulty, decision satisfaction, the perceived similarity and diversity of the options, or if there were enough choices. The questionnaire design was based on existing frameworks for user-centric evaluations of recommender systems [7, 11]. Furthermore, we made objective measurements including, e.g., the position of the selected item (list, index in list), the task completion time, or the number of items the participants inspected (item hover actions).

Table 1: UI Designs. Type: S=single list, M=multiple list

#	Algorithm to find similar items	Type
1	Content features	S
2	Tag-based similarity	S
3	Latent-vector similarity (SVD)	S
4	Content features (like #1, but ranked by SVD)	S
5	Content features (six features)	M
6	Best content features + tags + SVD	M
7	Like #6, but without duplicates	M
8	SVD (like #3, but presented with vacuous labels)	M

Participants. We recruited crowdworkers as participants through Amazon Mechanical Turk. For each condition, about 100 subjects participated. To ensure the reliability of the study, we only considered crowdworkers with a positive history (e.g., with more than 500 successful tasks performed in the past). Furthermore, we included an attention check in the post-task questionnaire to filter out inattentive crowdworkers. On average, the participants needed 5.25 minutes for the task for which they received a $1 payment. From 811 subjects, 775 passed the attention check, which makes us confident in the reliability of the crowdworkers.

Gender distribution among participants was almost balanced, 52% of the participants were male. The average age of the participants was about 39. Almost 45% declared that they use online movie services on a daily basis and, on average, the participants reported watching about 2-3 movies per week.

4 RESEARCH QUESTIONS AND OUTCOMES

The UI designs in Table 1 were chosen to explore alternative ways of presenting similar item recommendations and to answer a set of research questions (RQ). In the following, we address them by contrasting individual designs (experiment conditions). Descriptive statistics including means and standard deviations for all dependent variables are provided in the online material. Here, we summarize the main insights. Note that in the following, we are usually not interested if there are differences across all conditions (UI designs), as they are quite diverse, but in contrasting pairs of designs based on the specific research questions. Hence, we use pairwise t-tests with α=0.05 to test these explicit hypothesis. Only for RQ1, discussed next, where we compare more than two conditions, we use ANOVA and a Tukey post-hoc test for the analysis.

4.1 RQ1 – Value of Collaborative Information

Before investigating list selection aspects, we investigated the *general value of collaborative information* (tags, ratings) for finding similar items, following [13–15]. The insights from this analysis should guide the selection of the lists for the multi-list interfaces. We therefore compared the single-list designs #1, #2, and #3, where the selection of items in design #1 is based on the similarity of content features and designs #2 and #3 rely on social information.

The results showed that the SVD-based retrieval algorithm #3 was favorable over the content-based approach #1 in several dimensions, with the tag-based ranking #2 taking a middle position. The differences between Conditions #1 with #3 were significant in terms of perceived choice difficulty ($p < 0.01$), similarity to the reference movie ($p < 0.01$), and the perception of the quality of the logical ordering of the movies ($p = 0.03$). When asked if they found more than one suitable option and if there were sufficient alternatives, participants in Condition #3 and #2 gave more positive answers on average ($p < 0.001$).

To investigate the importance of rankings further, we designed Condition #4, where the selection of items is based on the content features (i.e., identical to Condition #1), but the items were ranked using the scores returned by the SVD-similarity method. We observed that the re-ranking was effective as participants picked items that appeared, on average, earlier in the list. They also found the presented items in general more similar to the reference list in Condition #4 than in Condition #1, although the same set of items was shown. At the same time, they found the recommendations in Condition #1 more diverse.

Overall, we conclude that social information should be considered when making similar-item recommendations in this domain, in particular when the goal is to reduce the perceived effort for participants.

4.2 RQ2 – Impact of List Optimization

Here, we investigate the *value of optimizing the set of lists to present* and specifically the value of including lists based on social signals. To this purpose, we first analyzed the behavior of participants in Condition #5, where they were presented six lists of recommendations based on content features. Note that the order of the lists in all multi-list conditions was randomized across participants. The labels used for the lists were chosen based on the underlying algorithm, e.g., "*Movies with the same genre*".

The analysis of the choice behavior in Condition #5 revealed that participants most frequently selected items from lists that showed movies from the same *genre* (29%) and *director* (26%). Items that were similar in terms of the *release date* and the *title* were the least frequently chosen (each about 5%). Suggestions based on *actor*-similarity, *image*-similarity, and *plot*-similarity took a middle position (9% to 13%).

Next, our aim was to investigate the effects of combining content features with social features, since social features were found particularly useful in previous studies like [13]. Therefore, in Condition #6, we replaced the rarely chosen lists based on release date and title with lists based on tags and ratings (SVD). Contrasting the content-based Condition #5 with the optimized list (Condition #6) revealed significant differences in certain dimensions. First, participants presented with the optimized list reported the task to be less difficult (mean response 2.32 vs. 3.06, p<0.001), they reported that it was easy to find more than one suitable item (5.75 vs. 5.27, p=0.02), and they considered the similarity with the reference items generally better (4.59 vs. 3.73, p<0.001). This speaks for the positive influence of optimizing the set of lists, in this case, by replacing the poorly-performing lists with ones based on social signals.

Overall, the analysis indicates that the optimization of the selection of lists to present may pay off in terms of user-perceived quality factors.

4.3 RQ3 – Effects of De-Duplication

Next, we were interested in whether de-duplication across multiple lists of recommendations matters. While de-duplication was mentioned as a processing step at Netflix in [5], in many other practical applications items can appear in more than one list.[4]

To analyze this question, we contrast Conditions #6 and #7. These multi-list conditions are identical, except that in Condition #6 we only retain the first occurrence of an item across multiple list. Remember that due to the randomization of the order of the lists, duplicate items may be removed from different lists for different participants.

Interestingly, the analysis did not reveal a significant difference in any of the dependent variables. We observed that the mean time needed for the choice increased from about 54 seconds to 65 seconds on average and that choice difficulty went up from 2.32 to 2.58. Such non-significant tendencies can be explained by the fact that more options are available. Overall, the absence of significant differences (p-values were between 0.15 and 0.30) in our study does of course not rule out that de-duplication may be meaningful in a given context.

4.4 RQ4 – Effects of Grouped Organization

The next research question addresses potential effects when the set of presented items is kept constant and only the visual arrangement is changed. To this end, we contrasted Condition #3, the SVD-based single-list UI design with Condition #8. In this latter condition, we took the 30 most similar items returned by the SVD method and randomly arranged them in six groups (lists) of five items. To avoid any biases that could originate from the labels for the lists, we used Greek letters as labels that carry no particular meaning.[5]

Our analysis revealed significant differences in a number of dimensions. First, when using the single-list UI, where the items are also strictly ordered by their SVD-based similarity, participants found the items more similar to the reference item and felt that they needed less time (even though the measured time revealed no significant difference). On the other hand, participants found the recommendations in the grouped organization on average more diverse (5.21 vs. 4.77, p=0.03) and novel (5.41 vs. 4.90, p=0.03), even though the exact same set of items was presented. Also, participants that used the single-list interface explored fewer options, measured in terms of item hover events (13.17 vs. 22.53, p<0.01).

Overall, we found that a grouped organization in multiple lists can contribute to creating the impression at the users' side of being provided with a set of suggestions that is both more diverse and novel. Also, the presentation of items in multiple lists and with an order that does not strictly follow the computed similarity, seemingly leads to more exploration.

5 DISCUSSION & RESEARCH LIMITATIONS

Discussion. On a general level and across the different analyses, our results indicate that single-list interfaces are favorable when it comes to the perceived effort of making a choice and the perceived similarity of the options with respect to the reference object. Interfaces with multiple lists, on the other hand, may lead to more exploration and to the impression of a more diverse and novel set of recommendations.

Which of these aspects is more desirable in the long run may in practice depend on the particularities of the application domain or even the business model of the provider. Quick decisions might for example lead to higher conversion rates and more sales in the short term, e.g., in e-commerce settings. In other domains, e.g., in the context of video or music streaming services, exploration, discovery, and user engagement in general are common key performance indicators, in particular when the service is based on a flat-rate subscription model.

At least in our experiments, the different UI designs did not lead to significant differences in terms of the participants' satisfaction with their decision in the short term. Conclusions about long-term effects are difficult to make. In practice, a longitudinal analysis and interpretation of the measured user interaction times may be required. Observing more item hover events and longer decision times can either mean that *(i)* users found many options they considered interesting or *(ii)* they had difficulties finding a suitable option. In practice, such aspects should therefore be monitored closely also over longer periods of time, and furthermore be correlated with other business-relevant performance metrics.

Research Limitations. While our exploratory study provided important insights for designers of recommender systems user interfaces, the presented study is so far limited to one particular domain, i.e., movie recommendations. Given previous research results from [13], we are confident that similar phenomena, e.g., regarding the value of social information, might exist in other domains of "quality-and-taste" as well, e.g., in recipe recommendation. However, it is unclear if our findings would translate to more objective domains, e.g., the recommendation of electronic devices.

Furthermore, our study so far was focused on one particular navigational situation, i.e., the presentation of similar objects with respect to a reference object. More research is therefore required to understand the effects of multi-list recommendation interfaces in

[4] Presenting the same item in a *single* list more than once was studied in [3]. No major effects of repeating items were however observed.

[5] Vacuous labels for different recommendation lists were previously used also in [4]. Biasing effects of different labels were also reported in [2] for a single-list recommendation interface.

different contexts, in particular on the landing page of the service or when search results are presented.

Another potential and more general limitation of user studies like ours is that participants interacted with an artificial system and that they might not be intrinsically motivated to accomplish the task and fill out the questionnaire. To assess the first risk (lack of realism) we included a corresponding question in the post-task questionnaire. With an average response of about 6 (on the 1-to-7 scale), we can consider the risk of lacking realism to be low. Regarding the lack of motivation, we found that over 95% of the participants passed the attention check in the post-task questionnaire. Since we only counted these participants, we are confident that the responses are sufficiently reliable.

6 SUMMARY

We presented the results of a first exploratory study on the use of multiple recommendation lists in the context of similar-item recommendations. The study suggests that single-list interfaces can be advantageous in terms of decision efficiency, but might at the same time lead to limited catalog exploration and discovery. With this work, we hope to stimulate more research in a so far largely underexplored area of high practical relevance.

ACKNOWLEDGMENTS

This research is funded by MediaFutures partners and the Research Council of Norway under grant number 309339.

REFERENCES

[1] Chris Alvino and Justin Basilico. 2015. Learning a Personalized Homepage. Netflix Tech Blog. Last accessed September 2020, https://netflixtechblog.com/learning-a-personalized-homepage-aa8ec670359a.

[2] Joeran Beel, Stefan Langer, and Marcel Genzmehr. 2013. Sponsored vs. Organic (Research Paper) Recommendations and the Impact of Labeling. In *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries*. 391–395.

[3] Joeran Beel, Stefan Langer, Marcel Genzmehr, and Andreas Nürnberger. 2013. Persistence in Recommender Systems: Giving the Same Recommendations to the Same Users Multiple Times. In *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries*. 386–390.

[4] Michael D. Ekstrand, Daniel Kluver, F. Maxwell Harper, and Joseph A. Konstan. 2015. Letting Users Choose Recommender Algorithms: An Experimental Study. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 11–18.

[5] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Transactions on Management Information Systems* 6, 4 (2015), 13:1–13:19.

[6] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders - Overview and Research Directions. *ACM Transactions on Intelligent Interactive Systems* 7, 3 (2017), 1–46.

[7] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22 (2012), 441–504.

[8] Pavlos Mitsoulis-Ntompos, Dionysios Varelas, Travis Brady, J. Eric Landry, Robert F. Dickerson, Timothy Renner, Chris Harris, Shuqin Ye, Abbas Amirabadi, Lisa Jones, and Javier Luis Cardo. 2020. Landing Page Personalization at Expedia Group. In *Proceedings of the 1st International Workshop on Industrial Recommendation Systems at KDD '20*.

[9] Ant Ozok, Quyin Fan, and Anthony F. Norcio. 2010. Design guidelines for effective recommender system interfaces based on a usability criteria conceptual model: Results from a college student population. *Behaviour & IT* 29 (2010), 57–83.

[10] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.

[11] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*. 157–164.

[12] Tobias Schnabel, Paul N. Bennett, and Thorsten Joachims. 2018. Improving Recommender Systems Beyond the Algorithm. *CoRR* abs/1802.07578 (2018).

[13] Christoph Trattner and Dietmar Jannach. 2019. Learning to Recommend Similar Items from Human Judgements. *User Modeling and User-Adapted Interaction* 30 (2019), 1–49.

[14] Yuan Yao and F. Maxwell Harper. 2018. Judging Similarity: A User-Centric Study of Related Item Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 288–296.

[15] Yifan Zhong, Tahir Lazaro Sousa Menezes, Vikas Kumar, Qian Zhao, and F. Maxwell Harper. 2018. A Field Study of Related Video Recommendations: Newest, Most Similar, or Most Relevant?. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 274–278.