User Perceptions of Diversity in Recommender Systems

PATRIK DOKOUPIL, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
LUDOVICO BORATTO, University of Cagliari, Italy
LADISLAV PESKA, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

In the context of recommender systems (RS), the concept of diversity is probably the most studied perspective beyond mere accuracy. Despite the extensive development of diversity measures and enhancement methods, the understanding of how users perceive diversity in recommendations remains limited. This gap hinders progress in multi-objective RS, as it challenges the alignment of algorithmic advancements with genuine user needs. Addressing this, our study delves into two key aspects of diversity perception in RS. We investigate user responses to recommendation lists generated using varied diversity metrics but identical diversification thresholds, and lists created with the same metrics but differing thresholds. Our findings reveal a user preference for metadata and content-based diversity metrics over collaborative ones. Interestingly, while users typically recognize more diversified lists as being more diverse in scenarios with significant diversification differences, this perception is not consistently linear and quickly diminishes when the diversification variance between lists is less pronounced. This study sheds light on the nuanced user perceptions of diversity in RS, providing valuable insights for the development of more user-centric recommendation algorithms. Study data and analysis scripts are available from https://bit.ly/diversity-perception.

CCS Concepts: • **Information systems** → **Recommender systems**.

Additional Key Words and Phrases: Diversity perception, Intra-list diversity, Binomial diversity, User study

ACM Reference Format:
Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2024. User Perceptions of Diversity in Recommender Systems. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), July 1–4, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3627043.3659555

1 INTRODUCTION

Motivation. In the evolving landscape of recommender systems (RS), the pursuit of accuracy has long been the main objective. Indeed, these systems have traditionally been evaluated by their ability to capture past user preferences and behaviors with high effectiveness [19]. However, the field of RS is witnessing a paradigm shift, with an additional focus on perspectives that go *beyond accuracy*. This shift has led to the emergence of Multi-Objective Recommender Systems (MORSs). These systems move from the traditional accuracy-centric framework, aiming to balance effectiveness with a broader spectrum of user-centric objectives [9, 16, 17, 31]. Among these objectives, *diversity* has emerged as a particularly important feature. Indeed, while accuracy ensures relevance, diversity enriches the user experience by enhancing user engagement, satisfaction, and potentially, discovery [10, 27].

Open issues. This focus on diversity has led to the development of a diverse array of metrics designed to inject variety into the recommendations [11, 13]. However, an essential piece of the puzzle remains underexplored, namely, *the user's perception of diversity in recommendation lists*. This aspect is crucial, as the success of a RS is ultimately measured not by objective metrics but rather by how users subjectively perceive and value the recommendations they receive.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2024 Copyright held by the owner/author(s).
Manuscript submitted to ACM

Understanding user perception of diversity is therefore central in aligning the technical advances in the development of MORSs with real-world user satisfaction and engagement.

Our contributions. To bridge this gap, in this paper, we present a study of how users perceive diversity in recommendation lists. We explore two primary axes: the diversity metric and the threshold indicating the intensity with which diversity is injected into the recommendation list. We analyze these dimensions by examining three distinct classes of diversity metrics: (i) *metadata*-based, where diversity is based on the attributes of the recommended items; (ii) *content*-based, where the intrinsic content of the items guides the diversification process; and (iii) *collaborative*, where diversity is derived from the variance in user ratings and preferences.

Specifically, in the following study, we aim to address three main research questions:

- RQ1: Having the same, fair, diversification procedure based on different diversity metrics, which metric users perceive as more diverse?
- RQ2: Having the same diversity metric, but using different diversification magnitudes, to what extent users perceive the increased diversity in the same way as indicated by the diversification thresholds?
- RQ3: Having user-perceived diversity results for recommended lists, which list-wise diversity metrics give the best estimation of the users' decisions?

To ensure that our findings have broad applicability and are generalizable, we conduct our study across two domains that have been widely studied in the RS literature: movie and book recommendation. Our results unveil interesting paradoxes in user perceptions. While users tend to prefer a genre-based diversity as providing more diversity under the same diversification conditions (RQ1), they are less capable of distinguishing between varying diversification thresholds of this specific metric (RQ2). Furthermore, despite the general tendency of users to evaluate cases with larger differences in diversification magnitude (RQ2), or with more agreeing metrics (RQ3) alike, there is still a large quantity of unobserved variance. Even for the most "clear" cases from the metrics point of view, users often disagree with the metrics estimations. This stresses the complexity of the problem and the need for proposing more user-aligned diversity metrics. Apart from addressing the main questions, we also provide additional insights into the differences between evaluated datasets and the impact of particular diversification procedures on respective diversity metrics.

1.1 Notations

Throughout the paper, we reserve lower-case u, v for users and i, j for items, while $\mathcal{U}$ and $\mathcal{I}$ denote the sets of all available users and items, respectively. We further denote the set of currently available candidate items as $C \subset \mathcal{I}$, the list of recommendations as L, existing user ratings as $r_{u,i} \in \mathbf{R}$, and estimated relevance as $\hat{r_{u,i}} \in \hat{\mathbf{R}}$. Diversity metrics are denoted as $d()$, where pair-wise $d(i, j)$ and list-wise $d(L)$ metrics can be distinguished based on the volume of parameters.

2 RELATED WORK

Following the pioneering work by Ziegler et al. [32], many variants of diversity definitions, diversity enhancement algorithms, and diversity evaluation metrics were proposed in the last two decades in the context of recommender systems. Due to space limitations, we do not intend to list them here but instead, refer interested readers to excellent overviews provided by Jesse et al. [11] (mainly focusing on intra-list diversity) and Kunaver and Požrl [13] (with a more general scope). However, despite the fact that diversity perception is intrinsic to the individuals, many of the proposed methods were not sufficiently grounded w.r.t. human cognition research or justified via experimental evaluation on real

users [11]. Nevertheless, there are some notable exceptions that focused on analyzing user perception of diversity (or similarity) in the context of recommender systems, e.g., [11, 22, 24, 27, 29]. In the remainder of this section, we will briefly discuss which of the related studies' design patterns we adopted and what are the key differences between ours and related studies.

First of all, we focus on the diversity perceived on the lists of recommendations. While this is also the case for some related works [11, 27], many others focused on comparing items in a *pair-wise* fashion, e.g., [22, 29], or focused on the list-wise *similarity* [24]. While results of pair-wise perceptions can be, to some extent, utilized to estimate list-wise features, we argue the mapping is imperfect in a similar fashion as, e.g., for pair- vs. list-wise learning to rank [6]. In [24], authors experimented on a different side of the similarity-diversity spectrum, i.e., whether delivering as similar results as possible will indeed be perceived as similar. Again, the generalizability of the results for more diversified lists is somewhat questionable. Also, instead of reporting the results on observing deployed 3rd-party websites as in [15], we opted for more freedom and controllability given by conducting a study in a mock-up environment. However, we strived to provide users with realistic stimuli and sufficient information needed for their decisions.

Let us now focus on the two most related studies [11, 27]. Similarly as in [11], we primarily focused on the intra-list-diversity-based metrics (with one addition), utilized rather smaller lists (eight instead of seven items due to display constraints), and employed a mixed design w.r.t. diversification magnitude (i.e., participants tried several, but not all combinations). Nonetheless, the main difference design-wise was the different construction of the displayed lists. In particular, [11] utilized static lists, while we generated them dynamically based on the users' responses in the preference elicitation phase. While the static design increases the controllability of the study, we argue that it might affect the practical relevance as the displayed recommendations might not align with what is normally displayed to users.

In this aspect, our study is more similar to [27]. Even though, the preference elicitation was enhanced to decrease its complexity (i.e., binary feedback only) and increase the chance of finding relevant items. Also, we utilized a similar diversification procedure as in Study 2 of [27]. In the study, we incorporated both purely relevance-based and purely diversity-based options, but instead of only one "mid-point" as in [27], we evaluated seven gradually increasing diversification magnitudes to provide more fine-grained results. On the other hand, we did not compare different list sizes, which was one of the main focuses of [27].

In contrast to both [11, 27], we adjusted the study flow as well. Instead of displaying lists sequentially and using a coarse-grained Likert scale to collect feedback, we displayed several lists simultaneously and provided users with a fairly fine-grained drag&drop UI (see Figure 2). This enabled us to observe multiple aspects of the user's judgment, i.e., absolute diversity perception, relative (pair-of-lists-wise) distance perception, and a binary comparison. We believe such a study design can provide valuable additional insights, outweighing the need to solve a more complex task. Also, we aimed at evaluating more variants of the diversity metrics, intentionally including those not tested in [11, 27] to increase the overall coverage of diversity metrics being evaluated under similar conditions. As for the study domains, we utilized movies (similarly as in [11, 27], but using a more up-to-date version of the MovieLens dataset: MovieLens Tag Genome 2021 [12]) and books (GoodBooks-10k dataset [30]), which is much less covered in general.

3 MODELS AND METHODS

3.1 Datasets and Pre-processing

The experimental evaluation was conducted on two domains: movies and books. In order to ensure realistic stimuli, we focused on larger, more recent datasets in both domains and further post-process them so that we could provide users

with enough data to evaluate the items. In particular, we utilized *MovieLens Tag Genome 2021* dataset [12] for the movies domain and *Goodbooks-10k* [30] for the books domain. Tag Genome 2021 dataset covers a broad selection of movies, $|\mathcal{I}| = 85K$, up till rather recent ones (newest from 2021). It also features links to external repositories (IMDb.com) and a sufficient volume of users and ratings to train a collaborative recommender ($|\mathcal{U}| = 247K$, $|\mathbf{R}| = 28.5M$). While the Goodbooks-10k dataset is slightly older (published in 2017), we argue that in the case of books, their recency has a lower impact on users. This dataset also contains a reasonable volume of data ($|\mathcal{U}| = 53K$, $|\mathcal{I}| = 10K$, $|\mathbf{R}| = 6.0M$) and links to external repositories. In order to collect additional information to display during the study (cover images, plot summaries, top book genres, etc.), we used data available from linked IMDb.com and GoodReads.com item profiles for movies and books, respectively. For the sake of brevity, the datasets will be further denoted as *MovieLens* and *GoodBooks*.

In order to fit their designated purpose, the datasets were further processed as follows. We filtered out all items for which a valid image or necessary metadata could not be obtained. In addition, for GoodBooks, we only kept a single top genre per book because otherwise, the computational complexity of the binomial diversity exceeds acceptable boundaries (i.e., more than a couple of seconds) while used in the diversification procedure. For MovieLens, we further filtered out very old movies (released before 1985), very old users (user's latest interaction is from 2016 or older), and then movies for which less than ten ratings remain (i.e., not recently used). After the pre-processing, the dataset sizes were $|\mathcal{U}| = 53K$, $|\mathcal{I}| = 6.5K$, $|\mathbf{R}| = 4.6M$ for GoodBooks and $|\mathcal{U}| = 18K$, $|\mathcal{I}| = 16K$, $|\mathbf{R}| = 5.1M$ for MovieLens.

3.2 Diversification Procedure

Over time, multiple diversification procedures were proposed to enhance the diversity of the lists of recommendations, mostly utilized in a recommendation's post-processing phase [1, 4, 17, 27, 32]. Some approaches are bound by particular definitions of diversity and, therefore, are not suitable for our use case, e.g., [21]. Out of the diversity-agnostic approaches, we selected a widely adopted bounded greedy optimization (see Algorithm 1) in a version close to [4] (with a few adaptations for the purposes of this study). At each step, the procedure selects the item with the highest marginal gain as the next item for the list. The gain is defined as a weighted average of the added relevance and added diversity, tuned by a hyperparameter α (higher values indicate more focus on diversity). The candidate options are limited (therefore, "bounded" greedy) as indicated by the candidate set C.

Algorithm 1 *Greedy_Div*: Bounded greedy diversification procedure.

input candidates $i \in C$, candidate's relevance $\hat{r}_i \in \hat{R}$, diversity metric $d(L)$, diversification magnitude α, list size k
$L \leftarrow []$
$\forall i \in C : \hat{r}_i^{norm} = eCDF_R(\hat{r}_i)$
for $c \in range(k)$ **do**
  $\forall i \in C : d_i^{norm} = eCDF_{d,L,C}(d(L + \{i\}) - d(L))$
  $i_{best} = argmax_{\forall i \in C}\left(\alpha * d_i^{norm} + (1 - \alpha) * \hat{r}_i^{norm}\right)$
  append i_{best} to L; remove i_{best} from C
end for
return L

Note that there are some modifications to the procedure variant as described in [4]. First, during our study, we aim to derive a fair comparison of different diversity definitions, which, however, may result in very different value distributions and, in turn, affect the diversification procedure. To alleviate this issue, we propose to transform both diversity and relevance marginal gains using an empirical cumulative distribution function (eCDF) trained on the

available data (i.e., relevances of candidate items R for relevance normalization and attainable diversity gains at each step for the diversity normalization). In this aspect, the procedure resembles a generalized version of the original ILD diversification procedure by Ziegler et al. [32].

Next, in accordance with the related work [4, 7, 10, 28], we decided to limit the volume of candidates $|C|$ so that the diversification procedure can be completed in a reasonable time.[1] Empirically, using around 500 candidates resulted in a satisfactory performance. However, if top-k candidates w.r.t. predicted relevance were supplied, we noticed an occasional issue with insufficient intrinsic diversity of the candidates (e.g., all blockbusters, all sharing some genres, etc.). As such, even for higher α values, we might receive insufficiently diverse lists because diverse items are simply not available in the candidates set. Therefore, we decided to select the top 250 candidates according to the estimated relevance, accompanied by another 250 items selected at random as the candidates set C.

3.3 Diversity metrics

Based on their underlying data, diversity metrics can be roughly clustered into the following groups: *collaborative*, *metadata-based*, and *content-based*. By *collaborative* diversity, we understand metrics that use users' feedback provided on individual items. This includes using raw feedback data (i.e., rows in the $\mathcal{R}$ matrix [2, 5]) or learned collaborative models (e.g., item embeddings of a trained matrix factorization [11, 20, 27]). By *metadata-based* diversity, we understand metrics that are calculated w.r.t. some (structured) attributes associated with the items. While the choice of appropriate metadata is inherently domain-specific (see, e.g., [10, 11, 25] for examples), using genres [11] or tags [24, 29] is among the prominent choices for multimedia domains.

Finally, by *content-based* diversity, we understand metrics that are based on the raw (unstructured) content data of the items. For multimedia domains (including books and movies), such metrics got in the research spotlight only recently with the advance of deep learning (DL) techniques. Depending on available data, content-based diversity metrics may include the similarities defined on top of video-, audio-, or text-based embeddings generated by appropriate DL techniques. Although the utilized datasets do not include raw content in its truest sense (i.e., full texts for books or video files for movies), we can use an approximation based on the free-text plot descriptions that are available for both datasets. One benefit of using textual data as compared to other multimedia is that meaningful free-text content is available in many domains (e.g., article texts in news, lyrics for music, or descriptions for e-commerce items, which makes the considered approaches broadly applicable).

For each class of diversity metrics, a plethora of variants were proposed, but to make this research feasible, we had to limit ourselves to only a few examples for each class. In general, we aimed at such definitions of diversity, that are applicable to a wider spectrum of domains and that are readily available in both datasets used in the study. In most considered cases, we relied on a broadly adopted intra-list diversity (ILD, [11, 32]), a meta-metric defining the list-wise diversity $d_{ILD}(L)$ as a mean of pairwise diversities $d(i, j)$ for all pairs of items in L.

$$d_{ILD}(L) = \frac{\sum_{i,j \in L; i \neq j} d(i, j)}{|L| * (|L| - 1)} \tag{1}$$

ILD's main advantage is that practically any pairwise metric can be incorporated into the equation (1), which makes it highly versatile. Also, its marginal gains (line 5 in Algorithm 1) can be easily re-defined, so that the pairwise diversities do not have to be repeatedly calculated. On the other hand, pairwise metrics might not fully reflect some important

[1] Especially MD-Genres-BinDiv variant (defined in Section 3.3) is rather computationally intensive.

aspects of diversity, e.g., size-awareness and non-redundancy, as shown by Vargas et al. [25]. Therefore, we also consider applying some of the more advanced metric definitions where suitable for the underlying data.

In particular, for collaborative diversities (further denoted as *CF-**), we considered two variants: (i) *CF-raw-ILD*, using ILD, where $d(i, j)$ is defined as a cosine distance of rating vectors corresponding to items i and j, and (ii) *CF-latent-ILD*, where $d(i, j)$ is defined as a cosine distance of row vectors corresponding to items i, j in the trained EASE model.[2] For metadata-based diversities (further denoted as *MD-**), we considered the following options: (i) *MD-tags-ILD*, using ILD, where $d(i, j)$ is defined as a cosine distance of tags associated with items i and j. In MovieLens, we utilized the TagGenome profile of individual items, while for GoodBooks, we utilized top-100 shelves corresponding to individual items as indicated in their respective GoodReads profiles and then post-process them to only include tags mentioned in at least 200 books. (ii) *MD-genres-ILD*, using ILD, where $d(i, j)$ is defined as a cosine distance of genres associated with items i and j. (iii) *MD-genres-BinDiv*, using binomial diversity [25] on top of genres assigned to the list members. Binomial diversity is defined as a product of two components: genres' *coverage* and their *non-redundancy* in the list of recommendations. Both components are defined as relative to the size of the list and estimated relevances of individual genres.[3] For content-based diversities (further denoted as *CB-**), we only considered one variant based on the plot summaries: (i) *CB-plot-ILD*. In particular, we applied a pre-trained CLIP [18] model[4] on the plot summaries of individual items to obtain their respective embeddings. For recommendation lists, the diversity is calculated as ILD with $d(i, j)$ corresponding to the cosine distance of respective embeddings. Similarly as in the case of CF-latent-ILD, we prioritized a more recent model over the classical alternatives used in the related work (e.g., TF-IDF and Latent Dirichlet Allocation in [22, 24]).[5]

3.3.1 Final Metrics Selection. Due to budget constraints, we needed to further shrink down the list of diversity metrics that can be directly utilized in the user study. Based on the study's power analysis, the maximal volume of metrics was set to three. In order to make the final cut, we decided to utilize the following selection criteria.

- **Conceptual variety:** Ideally, keep a representative from each of CF, MD, and CB groups.
- **Novelty & coverage:** Prefer metrics that are innovative or that were less experimented with recently.
- **Performance variety:** Do not include metrics that behave too similarly in a pre-study off-line evaluation.
- **Sanity checks:** Only include the metric if it provides a reasonable ordering for (manually checked) test cases.

For the off-line evaluation of the performance diversity, we randomly selected 1000 users from both MovieLens and GoodBooks datasets and predicted top-10 recommended items using the EASE algorithm. Then, each list was evaluated using all diversity metrics, while the per-metric results were compared via Pearson's correlation. Results of the off-line pre-study are depicted in Figure 1. For the sanity checks, we selected several well-known items from both domains and manually checked, which items are k-th closest to the source items ($k \in \{1, 10, 100, 1000, 5000\}$) w.r.t. individual metrics. Specifically, we evaluated whether the closest items are indeed similar and whether we can perceive the decreasing similarity trend for increasing k.

Considering the conceptual variety and novelty as our main selection criteria, we decided to select *CB-plot-ILD* as the only content-based candidate and *MD-genres-BinDiv* as, in contrast to the other alternatives, we are not aware of

[2]Note that in contrast to some related works (e.g., [27]), we did not use latent factors based on the matrix factorization, but rather this more up-to-date model, readily available for the study.
[3]Note that for genres' relevance, we only consider the global component (i.e., $\alpha = 0$ in eq. 8 of [25]).
[4]We utilized an OpenCLIP implementation, model 'ViT-B-32', pre-trained on 'laion2b_s34b_b79k' dataset (https://huggingface.co/laion/CLIP-ViT-B-32-laion2B-s34B-b79K).
[5]Note that CLIP was trained to provide a joint latent space for both texts and images and is considered one of the top models for text-based image search [14]. As such, it may also be a suitable backbone for heterogeneous diversity models (e.g., to combine textual and visual information).

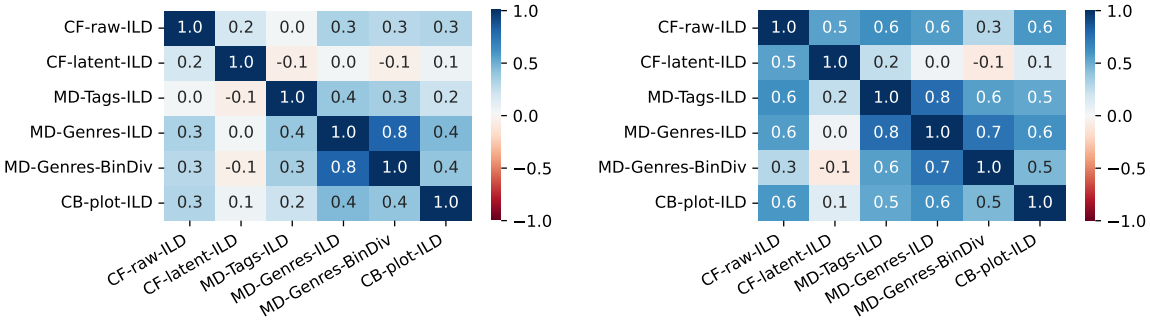

Fig. 1. Correlations between individual diversity metrics in off-line evaluation on MovieLens (left) and GoodBooks (right) datasets.

any recent studies evaluating the perception of genre-based binomial diversity metric. The decision was also supported by the fact that both metrics were not heavily correlated and passed the sanity checks. For collaborative variants, although *CF-latent-ILD* is less correlated with other selected approaches and, to the best of our knowledge, EASE-based embeddings were not evaluated as the basis for diversity yet, we were a little worried by the results of the sanity checks. Adding the fact that several recent user studies [11, 24] opted for a latent variant of collaborative diversity, we finally decided to include *CF-raw-ILD* so that the aggregated variety of the recent studies increases.

Note that despite only three metrics were selected to be directly utilized during the study, we calculated the scores for all considered metrics and all lists displayed to the users post hoc (see Section 5.3). Also note that in results (Tables 1, 2, and 4), metric names were abbreviated for the sake of space. We utilize *CF*, *MD*, and *CB* for the three selected metrics, and when referring to all considered metrics, the "*-ILD*" suffix was removed.

4 EXPERIMENTAL DESIGN

4.1 Study Flow

The study was conducted on the movies and books domains using the EasyStudy framework [8]. The user study was organized in six phases: *pre-study questionnaire*, *preference elicitation*, *diversity metric selection*, *diversification magnitude assessments*, *recommendations and post-study questionnaire* (not covered in this paper).

Pre-study. Prior to the study commencement, users were shown a study mission statement and detailed instructions and were asked for informed consent on the publication of anonymized data. Then, participants were routed to the pre-study questionnaire focusing on their familiarity with the domain, knowledge of recommender systems, and their objective criteria, as well as what corresponds to their perception of recommendations diversity. Exact formulations are available from the study repository: https://bit.ly/diversity-perception.

Preference Elicitation. In order to provide realistic recommendation lists in the subsequent phases, we first needed to solicit the preferences of users. To do so, we provided users with a list of items, asking them to indicate which items they had already experienced and liked (i.e., positive-only feedback). The prompts were sampled so that both popular and less known items were represented, and a sufficiently diverse selection was provided. In particular, we sampled eight items based on each of the following three criteria: popularity, novelty, and diversity. For popularity, we observed the mean rating of items in the training data (missing ratings counted as zeros), linearly scaled them to form a probability distribution over all items, and then sampled w.r.t. this distribution. The same procedure was applied for

Fig. 2. Example of the diversity magnitude assessment step. The top part depicts three diversified lists constructed using different α thresholds, while the bottom part contains drag&drop GUI to indicate diversity levels.

novelty as well, but in this case, using familiarity complement, i.e., $nov_i = 1 - |u : r_{i,u} \text{ exists}|/|\mathcal{U}|$ to generate probability distribution. For diversity, we considered items sampled in the previous bins and calculated the marginal gains of all remaining items w.r.t. CF-ILD [3]. Then, we applied the same procedure as above using this statistic.

Note that we only allowed users to proceed to the next steps after they selected at least five items. To make such a request realistic, users could always click on the "Load more data" button to generate additional 24 items or to manually search for particular items using a search bar.

Diversity Metric Selection. In this step, we focused on RQ1: given the same, fair conditions (i.e., the same *Greedy_Div* diversification procedure using the same diversification magnitude α), which results are perceived as more diverse? In particular, we displayed three lists of 8 recommendations to the users, asking them to indicate which of these lists appear as most diverse to them.[6] The lists were constructed using *Greedy_Div* diversification procedure (Algorithm 1) with a fixed diversification magnitude threshold $\alpha = 0.1$. Estimated relevance scores were predicted using the EASE algorithm [23] pre-trained on the source dataset (MovieLens or GoodBooks), where users were represented by their positive feedback indicated during the preference elicitation. Each list was constructed using a different diversity metric (i.e., *CF-raw-ILD*, *CB-plots-ILD*, or *MD-genres-BinDiv*). Note that the order of the lists was randomized, and we did not disclose to the users how they were constructed. The required users' feedback was to simply click on the list with the highest perceived diversity.

[6]The exact prompt was "Select a recommendation list that you perceive to be the **most diverse**."

Fig. 3. Information displayed upon hovering over an item for MovieLens (left) and GoodBooks (right) variants of the study.

Diversity Magnitude Assessment. Upon completing the metric selection step, users were routed towards two iterations of diversity magnitude assessment. In this step, we focused on RQ2: given the same diversification procedure and diversity metric, can the users perceive the differences in the outputs of different diversification magnitudes?

For each user, only one of the diversity metrics was considered. The metric assignment procedure was randomized, primarily aiming at maintaining similar volumes of users for each metric and secondly trying to assign the metric the users chose in the previous step frequently enough so that the results are balanced both w.r.t. metric variant and w.r.t. following user's preferred choices or not.

For the selected metric, users received two times three lists of recommendations (denoted as List 1 - 6) generated using the diversification procedure *Greedy_Div* using different diversification thresholds $\alpha \in \{0.0, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99, 1.0\}$ selected at random. The lists were displayed to the user side-by-side similarly as in the previous step. Nonetheless, this time, users were asked to utilize a drag & drop UI (see Figure 2) to specify how diverse they perceive the lists to be.[7] We utilized both the "face values" of the perceived diversity scores (i.e., absolute positions of the dropped objects), as well as their relative ordering in the data analysis (Section 5).

Recommendation Iterations. After the diversity magnitude assessments, users experienced several iterations of recommendations using multi-objective as well as relevance-only algorithms. These were interleaved with questionnaires to assess the quality of the provided recommendations. Although this section of the study was not analyzed in this paper, we need to mention that the questionnaires contained several attention checks based on which a portion of participants was rejected (if failing two or more checks).

4.2 Stimuli

During the study, users were confronted with lists of eight items (books or movies) and inquired about the diversity of depicted lists. At each study step, three lists were shown next to each other with clearly visible borders, and two items were displayed at each row for each list (see Figure 2). Minimal screen resolution requirements were imposed on participating users so that all items were visible to the user at once. Each item is represented by its cover image, while additional information was displayed while hovering over the image (see Figure 3). Also note that while "basic" instructions were always visible for the users at each step, we further provided a pop-up sidebar with more details available upon clicking on the "Instructions" button.

[7]The exact prompt was: "Use the drag&drop to order the recommendation lists from least diverse to most diverse."

4.3 Study Design Limitations

We consider as the main limitations of the study design the compromises that were required to simultaneously maintain a reasonable study duration, an affordable volume of participants, and reasonable study power. Therefore, only one diversification procedure and three instances of diversity metrics were selected to be evaluated with participants, diversification magnitudes were compared using a mixed design, and only a single $\alpha = 0.1$ threshold was used for metric assessments. While we strived to make logical and well-justified selections of the evaluated options, these still limit the generalizability of the results. To this end, we deliberately opted for including the second dataset, rather than including more diversity metric variants, as additional list-wise diversities can be calculated post hoc as shown in Section 5.3.

A similar limitation comes with a somewhat coarse selection of initially considered diversity metrics, which was imposed mainly for the sake of manageability. With one exception, we only considered one metric per information type and we only utilized simple ILD metrics. The exception was a genre-based binomial diversity, which is a theoretically well-justified metric using readily available data, but, to the best of our knowledge, not yet evaluated in a similar user study. Therefore, we opted to include it despite compromising the regularity of the study design a bit. Note, however, that finer-grained metric variants (e.g., focusing on additional data modalities, or different processing of current data, such as ranking-aware diversity [26]) can be evaluated post hoc, using the raw study results that were made available.

Finally, in this study, we solely focus on the perception of list-wise diversity by users. While, e.g., the questions related to the perceived relevance, relevance-diversity interplay, or pair-wise vs. list-wise diversities are undoubtedly important to answer, we did not want to over-complicate the study and possibly introduce unwanted biases by nudging users towards a particular way of thinking while assessing the list diversity.

5 RESULTS

The study was conducted in January 2024. In total, 268 participants were recruited using the Prolific.com service as well as via additional calls-for-participation sent out by the authors of the study. Prolific participants were pre-screened for fluent English, no less than ten previous submissions, and a 99% approval rate. Additional participants were recruited from the pool of users who participated in the author's previous studies and were not previously rejected due to the failed attention checks. Out of the recruited participants, 29 did not finish the study, while we further removed 23 participants who failed the attention checks, resulting in 216 valid participants (111 for the MovieLens and 105 for the GoodBooks, respectively). Usually, it took approx. 25 minutes to complete the whole study.

In the pre-study questionnaire, 83% of the users indicated having at least a little knowledge of recommender systems, while only two participants self-identified as experts in the area. Furthermore, 72% of users indicated that they had previously heard about the term "diversity" in the context of recommender systems. However, when inquired about what definition of diversity seems most appropriate for them, users frequently (in 33%) selected the prompt corresponding to the exploration criterion.[8] Note that we checked whether this subgroup behaved differently in the rest of the study. While these users tend to more often select *CB-plot-ILD* as the most diverse metric (see Section 5.1), the differences were not stat. sign., and no other notable differences were observed.

The rest of the section is organized according to the main research questions (RQ1-3).

[8]In particular, for the question "Please select the statement that best describes what diverse recommendations mean to you", 59% of users selected "Showing you a selection of very heterogeneous items (e.g., each one having a different set of genres).", 33% of users selected "Providing items that are highly different from the ones you consumed so far (e.g., genres you never watched so far).", while the remaining 8% selected prompts corresponding to novelty, exploitation, or a free-text answer.

Table 1. Overall results of diversity metric selection: Volume of selections per metric (left) and the ratio of concordant estimations based on list-wise metric values (right). Best results are in bold, while significant differences (p-val < 0.05) are denoted with an asterisk (*).

	Selected variant			$Mean\left(d(L_{selected}) > d(L_{not\_selected})\right)$						
Dataset	CF	MD	CB	CF-raw	CF-latent	MD-tags	MD-gen.	MD-BinDiv	CB-plots	1-Rel.
MovieLens	*20	**55**	36	*0.44	0.53	0.56	**0.64**	0.61	*0.49	0.58
GoodBooks	*23	**46**	36	*0.47	*0.50	0.60	0.58	0.59	*0.50	**0.65**
Overall	*43	**101**	*72	*0.45	*0.51	0.58	0.61	0.60	*0.50	**0.61**

5.1 RQ1: Using the same diversification procedure, which metric is evaluated as the most diverse?

To answer this research question, we analyzed the results of the diversity metric selection part of the study. As indicated in Table 1, the lists generated using *MD-genre-BinDiv* metric were selected the most often, followed by *CB-plots-ILD* and *CF-raw-ILD* attracted the least selections. To check the significance of observed data, we utilized a Chi^2 test with a post hoc pairwise comparison using Bonferroni correction. For individual datasets, we recorded significant differences *MD-genre-BinDiv* > *CF-raw-ILD* for GoodReads and (*MD-genre-BinDiv*, *CB-plot-ILD*) > *CF-raw-ILD* for MovieLens. If both datasets were aggregated, all pairwise relations were significant. We can conclude that when using the outputs of the same diversification procedure ($Greedy\_Div$) powered by different metrics, then already small diversification magnitude ($\alpha = 0.1$) provides perceivable differences among outputs of different metrics and *MD-genres-BinDiv* results are most often evaluated as the most diverse.

Note that we also checked, to what extent the selection might have been predicted based on the calculated diversities of the lists. Table 1 also contains the ratio of concordant pairs (i.e., diversity of the selected list is higher than the diversity of the not selected list) for all originally considered diversity metrics as well as w.r.t. an inverse to the mean normalized estimated relevance of the list. Out of the metrics present in the diversification procedure, *MD-genres-BinDiv* was indeed significantly more concordant with the user's choices than both other variants. Nonetheless, somewhat worrying was that an even better predictor of the outcome was the mean normalized relevance $\hat{r}_{u,i}^{norm}$ of the list (denoted as *1-Rel.* in the table, i.e., less relevant lists having a higher chance of being selected as the most diverse ones). We checked to what extent the observed differences can be explained by the differences in the per-list relevance scores (i.e., *MD-genres-ILD* lists would be considered as most diverse simply because they were less relevant), but this was not the case.[9] We hypothesize that, for low diversification magnitudes, users' perception of diversity may, to some extent, overlap with the perception of decreased relevance. Nonetheless, a dedicated future study is needed to verify this.

5.2 RQ2: To what extent are differences in diversification magnitude perceivable?

To answer the second research question, we analyzed the results of the diversity magnitude assessment part of the study. We considered three types of human judgments that can be derived from the drag&drop usage: (i) *absolute judgment*, i.e., using the dropped position for each list as its perceived diversity, (ii) *relative judgment*, i.e., using the difference between the positions of two lists as the perceived difference in the diversity magnitudes, and (iii) *relative binary judgment*, which merely consider the ordering of the list pairs, (i.e., binary information whether more diversified lists were indeed placed as more diverse by users). Note that for the relative variants, we only compared lists that were simultaneously depicted to the user (i.e., on the same page).

[9]The mean per-list relevance scores were 0.9991, 0.9989, and 0.9989 for *CF-raw-ILD*, *MD-genres-BinDiv*, and *CB-plot-ILD*, respectively, with no significant pairwise differences.

Table 2. Overall results of diversification magnitude perception. For absolute and relative judgments, we report Pearson's correlation, while for binary judgments, we report the ratio of concordant pairs. For each judgment type, the "ALL" column provides overall results, while "CF", "MD", and "CB" columns depict separate statistics for lists generated using CF-raw-ILD, MD-genres-BinDiv, and CB-plot-ILD, respectively. Best results are in bold, while significantly inferior (p-val < 0.05) results are denoted with an asterisk (*).

	Absolute judgments				Relative judgments				Relative binary			
Dataset	ALL	CF	MD	CB	ALL	CF	MD	CB	ALL	CF	MD	CB
MovieLens	0.26	0.25	0.26	**0.27**	0.30	**0.31**	0.28	0.30	0.60	**0.62**	0.59	0.59
GoodBooks	0.24	**0.35**	*0.11	*0.27	0.30	**0.41**	*0.15	0.35	0.63	0.65	0.58	**0.68**
Overall	0.25	**0.30**	*0.18	0.27	0.30	**0.36**	*0.21	0.32	0.62	**0.63**	0.58	0.63

Table 3. Comparing ratios of concordant pairs for different levels of diversification magnitude. Results significantly lower (p-value < 0.05 w.r.t. Fisher's exact test) than the highest ones are marked with an asterisk (*). Results significantly higher than the lowest ones are denoted with a circle (∘).

	Cluster					
Dataset	LOW	MIDDLE	HIGH	LOW-MIDDLE	LOW-HIGH	MIDDLE-HIGH
MovieLens	* 0.57	* 0.56	* *0.53*	* 0.62	∘ **0.73**	* 0.57
GoodBooks	∘ **0.72**	* *0.42*	∘ 0.66	∘ 0.69	∘ 0.67	∘ 0.66
Overall	∘ 0.64	* *0.50*	* 0.59	∘ 0.66	∘ **0.70**	*∘ 0.61

Table 2 contains the results for all three judgment types, individual datasets, and also separately for the source diversity metric utilized in the diversification procedure. Results reveal an interesting tradeoff: although users more often select *MD-genres-BinDiv* as the most diverse in the metric assignment step, they are less capable of distinguishing the magnitude of diversification w.r.t. this metric than for both of its counterparts. The effect is significant[10] for both absolute and relative judgments in the GoodReads dataset and prevails in the overall results as well. The differences w.r.t. binary judgments as well as for the MovieLens dataset were, however, not significant.

We further focused on differences in the metric-user agreements w.r.t. particular values of the diversification thresholds. To do so, we jointly denote the following clusters: $\alpha \in \{0.0, 0.01, 0.1\}$ as LOW diversification, $\alpha \in \{0.25, 0.5, 0.75\}$ as MEDIUM diversification and $\alpha \in \{0.9, 0.99, 1.0\}$ as HIGH diversification. The results, as outlined in Table 3, highlight that users were more aligned with the diversification procedure when ordering pairs of lists from different clusters (i.e., with higher differences in diversification magnitude), as compared to the lists from the same cluster (0.65 vs 0.57, p-value < 0.01). While the MEDIUM cluster seemed the most challenging for users, its comparison against HIGH lacked statistical significance. Dataset-specific analysis revealed some small variations, e.g., the easiest to judge were lists from the LOW cluster in GoodBooks, while LOW-HIGH cluster performed better in the MovieLens variant.

Finally, we also checked to what extent the simultaneous presence of lists (thus supporting direct relative comparisons of lists) had an effect on users' pairwise judgments. To do so, we compared the relative judgments between simultaneously displayed lists (i.e., displayed in the same step) vs. relative judgments between sequentially displayed lists (i.e., one displayed in the first step, while the other in the second). Indeed, we observed a significant drop in the overall correlations (0.30 vs. 0.26), suggesting that a direct comparison nudges users to report more consistently to the diversification magnitude. The effect is similarly strong in both datasets and affects all diversification procedure variants (more substantially *CF-raw-ILD* and *CB-plot-ILD*).

[10]Using z-test on Fisher's z-transformation for correlation coefficients and p-val < 0.05.

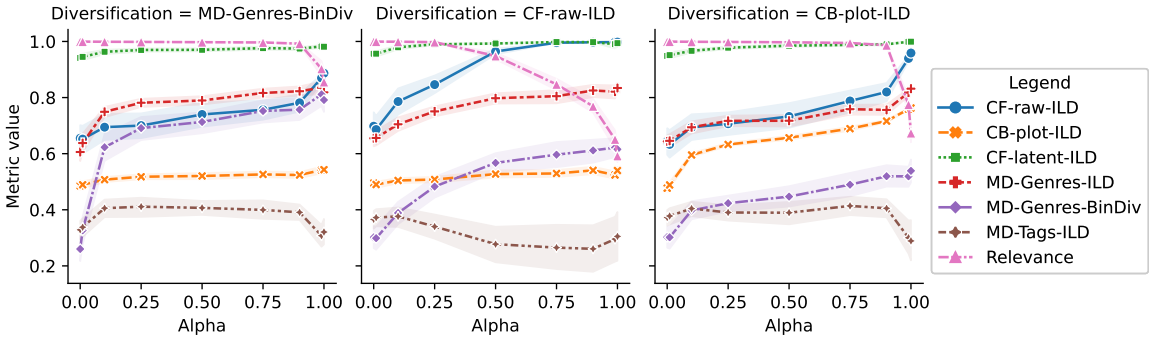

Fig. 4. Impact of individual diversification procedures on diversity and relevance values of generated lists.

Table 4. Comparing user's diversity perception vs. list-wise diversity metrics. Best results are in bold, while significantly inferior (p-val < 0.05) results are denoted with an asterisk (*).

		CF-raw	CF-latent	MD-tags	MD-gen.	MD-BinDiv	CB-plots	1-Rel.
Relative judgments	MovieLens	0.27	*0.14	*-0.16	0.22	0.27	**0.28**	*0.18
	GoodBooks	0.39	0.36	0.38	**0.41**	0.38	*0.28	*0.26
	Overall	0.28	0.26	*0.0	0.32	**0.33**	0.28	*0.21
Binary judgments	MovieLens	**0.61**	0.56	*0.41	0.57	0.56	0.6	0.6
	GoodBooks	0.61	**0.63**	0.61	0.6	0.6	0.57	0.62
	Overall	0.61	0.59	*0.51	0.58	0.58	0.58	**0.61**

5.3 RQ3: To what extent can per-list diversity values explain the perceived diversity?

To answer this research question, we post hoc calculated diversity values for all lists that were displayed to the users and w.r.t. all considered metrics. Then, we focused on the same judgment types as in RQ2. First, let us comment on how the diversification procedure affects the diversity values and how this might have impacted the following analysis. Figure 4 displays the mean and 95% confidence interval errors for each per-list metric aggregated w.r.t. diversification procedure and its magnitude. The figure clearly indicates that while using small diversification magnitudes ($\alpha \leq 0.1$), most of the considered diversification metrics are jointly increased to some extent. Nonetheless, higher diversification thresholds tend to provide diminishing or even negative returns for most metrics and substantially improve only the target one. The decrease in mean normalized relevance is very small for $\alpha \leq 0.75$ while using *MD-genres-BinDiv*- and *CB-plot-ILD*-based diversification. For *CF-raw-ILD*, the drop in relevance is quicker, which may be a natural effect of highly contradicting objectives. These results somewhat decrease the importance of diversity metric selection, as long as only minor diversification adjustments are planned.

The above-mentioned results pose one important limitation for the subsequent analysis. Despite some of the metrics that were not directly utilized for diversification during the study appearing to reflect well the user-perceived diversity, we cannot guarantee that the same result will be achieved if the lists were diversified using this metric. Nevertheless, considering the diversification procedure as black-box, we believe the following results can still provide some valuable insights and serve as a benchmark for subsequent analyses. Table 4 depicts the results for relative and binary judgments, while we omit the absolute ones for the sake of space. One can notice a rather stable good performance of *CF-raw-ILD* for all datasets and both judgment types, which is in line with the results of RQ2. The performance of both *CF-latent-ILD*

and *MD-tags-ILD* was inferior by a large margin in the MovieLens dataset for relative judgments, but rather good in the GoodReads dataset. While in the case of *MD-tags-ILD*, we may argue with possibly varying underlying information quality between datasets, this is not the case with *CF-latent-ILD*. Therefore, we would recommend being cautious while using latent collaborative models for diversity estimation. Overall, we noticed rather substantial differences between the datasets in terms of both overall values and the suitability of individual metrics, making inter-domain generalization rather tricky. This is in line with the observations of [11], who reported on major differences in the effectivity of different ILD metrics in movies and recipe domains. Finally, while the results of the relative judgments mostly correspond to the binary ones, the one notable exception was estimated relevance. While it may be used as a rather good estimator of users' binary judgments, its correlation with users' relative judgments is rather low.

Finally, we evaluated the potential of combining multiple metrics together to improve the predictions. To do so, we devised a simple majority-vote-based ensemble model out of all diversity metrics plus the inverse relevance to predict users' binary judgments. This ensemble prediction correctly estimated users' decisions in 61% of cases, i.e., it is similarly good as the best single method. Nonetheless, the estimation quality depends greatly on the number of agreeing models, i.e., 51%, 56%, 65%, and 72% for 4, 5, 6, and all 7 agreeing methods, respectively. These results show an additive power of the diversity metrics to distinguish between more and less clear cases for the users but also reveal a substantial knowledge gap that prevails in the data. Even if all considered metrics agreed with each other, users evaluated the pair of lists differently in 28% of cases.

6 CONCLUSIONS AND FUTURE WORK

Recommendation diversity has long been considered a central property when moving to properties beyond the accuracy of the results. Despite its significance, there has been a notable gap in understanding how users actually perceive and value diversity in recommendation lists. This paper aimed to bridge this gap by conducting a user study, delving into user perception of diversity in recommender systems. We examined how users perceive different diversity metrics and diversification thresholds. The results of our study revealed several key insights. Firstly, users showed a preference for metadata-based diversity metrics, suggesting that users associate diversity more with content types or genres. Secondly, while users could recognize increased diversity in scenarios with significant diversification differences, this perception was not always linear and tended to diminish in scenarios with subtler diversification variances. Thirdly, the study highlighted the complex interplay between perceived diversity and relevance, indicating that users might sometimes perceive less relevant lists as more diverse. Lastly, the effectiveness of diversity metrics in theory did not always align with perceived diversity in practice, emphasizing the need for user-centric evaluations of these metrics.

Looking ahead, several avenues for future work emerge from our study. There is a clear need for further research into understanding the optimal balance between diversity and relevance in recommender systems. Future work also will explore how different presentation formats and contexts influence user perception of diversity. Additionally, understanding the underlying factors that influence user perception of diversity and tailoring recommender systems to it remains a significant challenge.

ACKNOWLEDGMENTS

This paper has been supported by Czech Science Foundation (GAČR) project 22-21696S, Charles University grant SVV-260698/2023, and Charles University Grant Agency (GA UK) project number 188322.

REFERENCES

[1] Adomavicius, G., Kwon, Y.: Improving aggregate recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering **24**(5), 896–911 (2012). https://doi.org/10.1109/TKDE.2011.15

[2] Aytekin, T., Karakaya, M.Ö.: Clustering-based diversity improvement in top-n recommendation. Journal of Intelligent Information Systems **42**(1), 1–18 (Feb 2014). https://doi.org/10.1007/s10844-013-0252-9, https://doi.org/10.1007/s10844-013-0252-9

[3] Bradley, K., Smyth, B.: Improving recommendation diversity. In: Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science, Maynooth, Ireland. vol. 85, pp. 141–152. Citeseer (2001)

[4] Bridge, D., Kelly, J.P.: Ways of computing diverse collaborative recommendations. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) Adaptive Hypermedia and Adaptive Web-Based Systems. pp. 41–50. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

[5] Bridge, D., Kelly, J.P.: Ways of computing diverse collaborative recommendations. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) Adaptive Hypermedia and Adaptive Web-Based Systems. pp. 41–50. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

[6] Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th International Conference on Machine Learning. p. 129–136. ICML '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1273496.1273513, https://doi.org/10.1145/1273496.1273513

[7] Di Noia, T., Ostuni, V.C., Rosati, J., Tomeo, P., Di Sciascio, E.: An analysis of users' propensity toward diversity in recommendations. In: Proceedings of the 8th ACM Conference on Recommender Systems. p. 285–288. RecSys '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2645710.2645774, https://doi.org/10.1145/2645710.2645774

[8] Dokoupil, P., Peska, L.: Easystudy: Framework for easy deployment of user studies on recommender systems. In: Proceedings of the 17th ACM Conference on Recommender Systems. p. 1196–1199. RecSys '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3604915.3610640, https://doi.org/10.1145/3604915.3610640

[9] Dokoupil, P., Peska, L., Boratto, L.: Looks can be deceiving: Linking user-item interactions and user's propensity towards multi-objective recommendations. In: Zhang, J., Chen, L., Berkovsky, S., Zhang, M., Noia, T.D., Basilico, J., Pizzato, L., Song, Y. (eds.) Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023. pp. 912–918. ACM (2023). https://doi.org/10.1145/3604915.3608848, https://doi.org/10.1145/3604915.3608848

[10] Du, Y., Ranwez, S., Sutton-Charani, N., Ranwez, V.: Is diversity optimization always suitable? toward a better understanding of diversity within recommendation approaches. Information Processing Management **58**(6), 102721 (2021). https://doi.org/https://doi.org/10.1016/j.ipm.2021.102721, https://www.sciencedirect.com/science/article/pii/S0306457321002053

[11] Jesse, M., Bauer, C., Jannach, D.: Intra-list similarity and human diversity perceptions of recommendations: the details matter. User Modeling and User-Adapted Interaction **33**(4), 769–802 (Sep 2023). https://doi.org/10.1007/s11257-022-09351-w, https://doi.org/10.1007/s11257-022-09351-w

[12] Kotkov, D., Maslov, A., Neovius, M.: Revisiting the tag relevance prediction problem. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1768–1772. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463019, https://doi.org/10.1145/3404835.3463019

[13] Kunaver, M., Požrl, T.: Diversity in recommender systems – a survey. Knowledge-Based Systems **123**, 154–162 (2017). https://doi.org/https://doi.org/10.1016/j.knosys.2017.02.009, https://www.sciencedirect.com/science/article/pii/S0950705117300680

[14] Lokoč, J., Andreadis, S., Bailer, W., Duane, A., Gurrin, C., Ma, Z., Messina, N., Nguyen, T.N., Peška, L., Rossetto, L., Sauter, L., Schall, K., Schoeffmann, K., Khan, O.S., Spiess, F., Vadicamo, L., Vrochidis, S.: Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th vbs. Multimedia Systems **29**(6), 3481–3504 (Dec 2023). https://doi.org/10.1007/s00530-023-01143-5, https://doi.org/10.1007/s00530-023-01143-5

[15] Nilashi, M., Jannach, D., bin Ibrahim, O., Esfahani, M.D., Ahmadi, H.: Recommendation quality, transparency, and website quality for trust-building in recommendation agents. Electronic Commerce Research and Applications **19**, 70–84 (2016). https://doi.org/https://doi.org/10.1016/j.elerap.2016.09.003, https://www.sciencedirect.com/science/article/pii/S1567422316300485

[16] Paparella, V., Anelli, V.W., Boratto, L., Noia, T.D.: Reproducibility of multi-objective reinforcement learning recommendation: Interplay between effectiveness and beyond-accuracy perspectives. In: Zhang, J., Chen, L., Berkovsky, S., Zhang, M., Noia, T.D., Basilico, J., Pizzato, L., Song, Y. (eds.) Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023. pp. 467–478. ACM (2023). https://doi.org/10.1145/3604915.3609493, https://doi.org/10.1145/3604915.3609493

[17] Peska, L., Dokoupil, P.: Towards results-level proportionality for multi-objective recommender systems. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1963–1968. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531787, https://doi.org/10.1145/3477495.3531787

[18] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html

[19] Ricci, F., Rokach, L., Shapira, B.: Recommender systems: Techniques, applications, and challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 1–35. Springer US (2022). https://doi.org/10.1007/978-1-0716-2197-4_1, https://doi.org/10.1007/978-1-0716-2197-4_1

[20] Shi, Y., Zhao, X., Wang, J., Larson, M., Hanjalic, A.: Adaptive diversification of recommendation results via latent factor portfolio. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 175–184. SIGIR '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2348283.2348310, https://doi.org/10.1145/2348283.2348310

[21] Slaney, M., White, W.: Measuring playlist diversity for recommendation systems. In: Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia. p. 77–82. AMCMM '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1178723.1178735, https://doi.org/10.1145/1178723.1178735

[22] Starke, A., Larsen, S., Trattner, C.: Predicting feature-based similarity in the news domain using human judgments. In: INRA'21: 9th International Workshop on News Recommendation and Analytics, September 25, 2021, Amsterdam, Netherlands. CEUR-WS (2021)

[23] Steck, H.: Embarrassingly shallow autoencoders for sparse data. In: The World Wide Web Conference. p. 3251–3257. WWW '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3308558.3313710, https://doi.org/10.1145/3308558.3313710

[24] Trattner, C., Jannach, D.: Learning to recommend similar items from human judgments. User Modeling and User-Adapted Interaction **30**(1), 1–49 (Mar 2020). https://doi.org/10.1007/s11257-019-09245-4, https://doi.org/10.1007/s11257-019-09245-4

[25] Vargas, S., Baltrunas, L., Karatzoglou, A., Castells, P.: Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: Proceedings of the 8th ACM Conference on Recommender Systems. p. 209–216. RecSys '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2645710.2645743, https://doi.org/10.1145/2645710.2645743

[26] Vargas, S., Castells, P.: Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. p. 109–116. RecSys '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2043932.2043955, https://doi.org/10.1145/2043932.2043955

[27] Willemsen, M.C., Graus, M.P., Knijnenburg, B.P.: Understanding the role of latent feature diversification on choice difficulty and satisfaction. User Modeling and User-Adapted Interaction **26**(4), 347–389 (Oct 2016). https://doi.org/10.1007/s11257-016-9178-6, https://doi.org/10.1007/s11257-016-9178-6

[28] Wu, W., Chen, L., Zhao, Y.: Personalizing recommendation diversity based on user personality. User Modeling and User-Adapted Interaction **28**(3), 237–276 (Aug 2018). https://doi.org/10.1007/s11257-018-9205-x, https://doi.org/10.1007/s11257-018-9205-x

[29] Yao, Y., Harper, F.M.: Judging similarity: a user-centric study of related item recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems. p. 288–296. RecSys '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3240323.3240351, https://doi.org/10.1145/3240323.3240351

[30] Zajac, Z.: Goodbooks-10k: a new dataset for book recommendations. FastML (2017)

[31] Zheng, Y., Wang, D.X.: A survey of recommender systems with multi-objective optimization. Neurocomputing **474**, 141–153 (2022). https://doi.org/10.1016/j.neucom.2021.11.041, https://doi.org/10.1016/j.neucom.2021.11.041

[32] Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web. p. 22–32. WWW '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1060745.1060754, https://doi.org/10.1145/1060745.1060754