ABSTRACT

When recommendations become increasingly personalized, users are often presented with a narrower range of content. To mitigate this issue, diversity-enhanced user interfaces for recommender systems have in the past found to be effective in increasing overall user satisfaction with recommendations. However, users may have different requirements for diversity, and consequently different visualization requirements. In this paper, we evaluate two visual user interfaces, SimBub and ComBub, to present the diversity of a music recommender system from different perspectives. SimBub is a baseline bubble chart that shows music genres and popularity by color and size, respectively. In addition, ComBub visualizes selected audio features along the X and Y axis in a more advanced and complex visualization. Our goal is to investigate how individual traits such as musical sophistication (MS) and visual memory (VM) influence the satisfaction of the visualization for perceived music diversity, overall usability, and support to identify blind-spots. We hypothesize that music experts, or people with better visual memory, will perceive higher diversity in ComBub than SimBub. A within-subjects user study (N=83) is conducted to compare these two visualizations. Results of our study show that participants with high MS and VM tend to perceive significantly higher diversity from ComBub compared to SimBub. In contrast, participants with low MS perceived significantly higher diversity from SimBub than ComBub; however, no significant result is found for the participants with low VM. Our research findings show the necessity of considering individual traits while designing diversity-aware interfaces.

CCS CONCEPTS

• Human-centered computing → Visualization design and evaluation methods; User models; • Information systems → Recommender systems;

KEYWORDS

Individual traits; diversity; recommender user interfaces; visual memory; musical sophistication

1 INTRODUCTION

Recommender systems are increasingly used in various on-line application domains, such as e-commerce and technology-enhanced learning. They provide users with personalized items based on a variety of information, including preferences, on-line history, and demographics. Traditionally, most researchers focused on the accuracy of recommender systems by proposing various algorithms to increase precision and recall. However, recent studies have shown that the overall user experience of recommender systems is affected by many factors beyond accuracy [15] in some domains, such as diversity and serendipity for the music recommenders. Recommender systems aim at helping users explore new items of interest [18]. Diversity is important because users will not be satisfied with recommendations that are overly similar to what they have consumed previously [19]. Moreover, diversity enables comparison among recommendations, thereby increasing the confidence in making a choice [8, 25].

In fact, the increase of personalization may prevent users from consuming diverse content, which is called the “filter bubble”. To mitigate the issue, many approaches propose a diversity-enhancing algorithm, e.g., topic diversification [45], and a hybrid way of using multiple sources [41]. Other work [20, 37, 42] empowers users to manipulate recommender systems through an interactive user interface to explore items in a different way. Although the ranked list interface is still one of the most common ways to present recommendation results, several limitations have been identified [11]. Users tend to for instance pay less attention to the items at the bottom of the list [7]. In addition, predicting the perception of the diversity of items in a list is difficult. Recently, some researchers have proposed diversity-aware interfaces that present recommendations in different kinds of visualizations, such as two-dimensional scatter plots [37]. When using such an interface, users tend to perceive a higher diversity of content than in a ranked list interface.

In this paper, we explore how individual traits influence the effectiveness of a diversity-enhanced user interface. The individual traits that we are researching are musical sophistication (MS), that measures the ability to engage with music effectively [28], and visual memory (VM), that is temporary storage and manipulation of the information over an extended period of time [5]. We employ a design-based approach to explore how MS and VM can influence the effectiveness of the user interface. We developed a music recommender based on Spotify API 1. We designed and implemented two user interfaces: SimBub is the simplest visualization that clusters items by genres and shows popularity in a bubble chart. ComBub is more advanced and complex, and presents recommendations in a bubble chart with genres, popularity, and two

1https://developer.spotify.com/web-api/get-recommendations/
we attempt to answer the following three research questions:

1. How will the individual traits of musical sophistication and visual memory capacity influence the perceived diversity of recommended songs?
2. In general, does visualizing the audio features of music increase perceived diversity of recommended songs?
3. Does a particular interface encourage users to explore blind-spots, which might mitigate the issue of the “filter bubble”?

The remainder of the paper is organized as follows: after providing an overview of related research work on user interfaces for recommender systems and also some diversity enhancing approaches, we describe the design of our system. Then we describe the study design, followed by the study results, limitations, and discussion. Finally, we present the conclusions and future work.

2 RELATED WORK

Our research refers to three important topics of recommender systems: 1) using visualization techniques to enhance the user experience of the recommender system, 2) techniques to enhance recommendation diversity, 3) effects of individual traits on perceived diversity. We will briefly review related work of these topics.

2.1 Visualizations for Recommender Systems

Visualizations for recommender systems have been increasingly researched to improve various aspects of recommender systems. By visualizing the user profiles and recommendation process, transparency and user control of the system can be improved significantly. Jin et al. [21] for instance demonstrate an interactive flow-chart based visualization that explains how a selected ad is filtered for the targeted user profile. Verbert et al. [39] present a system that increases the effectiveness of making a choice by explaining the provenance of recommendations and offering control to users. Some systems show increased accuracy by enabling users to adjust the weight of algorithm parameters [6, 31]. In addition, several studies show the positive effects of visualization on perceived diversity [20, 33, 37, 42]. Our study focuses on similar diversity-aware visualizations. More specifically, we investigate the effects of individual traits on perceived diversity through different visualizations.

2.2 Individual Traits for Perceived Diversity

In fact, the desired level of diversity for a satisfying recommender differs among users [3], which may be attributed to different preferences for agreeable and challenging items [29]. Therefore, it is beneficial to adapt diversification strategies to individual traits.

Several research works [12, 13, 35] have shown that visual memory has a significant impact on visualization effectiveness. Moreover, in the framework for the user-centric evaluation of recommender systems [25], a significantly positive effect of expertise (personal characteristics) on diversity (subjective system aspects) has been found. Music experts may, for instance, perceive higher diversity than laymen from the same recommendation list. Moreover, personality is also an important factor to perceived diversity. In our study, we measure a user’s domain knowledge using musical sophistication index (MSI).

In the literature, we find two individual traits that may have an impact on the effectiveness of diversity-aware interfaces: musical sophistication (MS) and visual memory (VM). In this paper, we hypothesize that individual traits such as musical sophistication and visual memory can influence users' perceived diversity.

selected audio features. In a music recommender, we hypothesize that ComBub, that presents additional audio features, may better support users with higher MS and VM to perceive diversity. On the contrary, the attributes of genres and popularity shown in SimBub may be sufficient to help users with low MS and VM to perceive diversity. In addition, we consider both item and user diversity for choosing a strategy to present recommendations by following a diversity-aware recommendation model (see Figure 1) [34]. More specifically, the user diversity is mainly measured by the user’s musical sophistication and visual memory, and item diversity can be calculated by genres, popularity, and audio features.

To verify our assumption, we conduct a within-subjects user study, comparing the two interfaces, SimBub and ComBub, with an identical recommender algorithm and recommender sources. We employed ResQue, a user-centric evaluation framework for recommender system [32], to measure perceived diversity and other key aspects. Meanwhile, we are also interested in knowing if a specific interface encourages users to look for songs in under-explored genres (blind-spots), which has been researched in [36]. From this study, we attempt to answer the following three research questions:

- **RQ1**: In general, does visualizing the audio features of music increase perceived diversity of recommended songs?
- **RQ2**: How will the individual traits of musical sophistication and visual memory capacity influence the perceived diversity of the support?
- **RQ3**: Does a particular interface encourage users to look for items in under-explored genres?

The contributions of this paper are three-fold.

1. We designed a diversity-aware user interface ComBub where users are able to explore diverse items from different aspects. Comparing to the baseline user interface SimBub, ComBub shows a great potential to augment user perceived diversity for the users with high MS and high VM.
2. By measuring user’s MS and VM before the study, we investigated the effects of the two individual traits on the perceived diversity of the interfaces. The results reveal a strong positive correlation between individual traits and the perceived diversity of the interface(s).
3. We also investigated to what extent each interface is able to encourage users to explore blind-spots, which might mitigate the issue of the “filter bubble”.

Figure 1: Diversity-aware recommendation model fitted to our research contexts.
visual memory tend to influence the perceived diversity through visual interfaces, which to the best of our knowledge has not been investigated yet.

2.3 Diversity-enhancing Approaches

The definition of diversity may vary in different application domains [9]. In this paper, diversity refers to the diversity of a recommended list measured by an intra-list similarity metric [45]. The adaptation of content to individual preference is normally associated with lower diversity of content [30]. To address this problem, researchers have proposed to enhance diversity, either with novel algorithms or user interfaces.

2.3.1 Diversity-aware Algorithms. Re-ranking is a widely used technique to enhance diversity by aggregating diversity and accuracy in the process of recommendations [1, 2, 38, 44]. Furthermore, Zhang et al. [43] present an optimization method to improve both diversity and accuracy for the top-N prediction problem. Ziegler et al. [45] propose an approach to diversifying content based on topics. Similarly, Giannopoulos et al. [16] present an algorithm that diversifies news according to opinion similarities. Several studies [10, 26] propose various methods to generate critiques for increasing diversity in a critiquing-based recommender system. To remedy the limitation of imposing diversity on an existing system, McGint et al. [27] demonstrate an adaptive diversity-enhancing algorithm in a conversational recommender system. Most of these algorithms improve diversity for a ranked list. However, from the perspective of the user, the increased diversity in a ranked list is difficult to predict due to the position bias of the list as explained earlier.

2.3.2 Diversity-aware User Interfaces. Comparing to algorithmic approaches, fewer works are found to enhance perceived diversity through the user interface. Graells-Garrido et al. [17] present the distance of latent topics in a visualization, which supports active exploration of diverse content by users. Schafer et al. [33] present a user interface with personalized control that allows the user to find more diverse content generated from multiple information sources. Faridani et al. [14] present a map-based interface that enables users to explore diverse on-line comments. Hu et al. [20] propose an organization based interface to increase users’ perceived diversity of recommendations. Wong et al. [42] present a system named Diversity Donut that allows users to indicate the level of diversity for the recommended items. Tsai et al. [37] present a diversity-enhanced interface that presents recommendations with multiple attributes in a two-dimensional scatter plot inspiring our approach.

These previous interfaces are designed without considering individual requirements for diversity. In our work, we design two diversity-aware user interfaces that show different item attributes. As a result, the study results show the necessity of tailoring the interface design to individual traits.

3 SYSTEM DESIGN

In this section, we first explain a seed-based recommendation algorithm implemented for a music recommender, and then we describe two interfaces designed for enhancing recommendation diversity.

3.1 Algorithm

The recommendation algorithm was implemented by leveraging the Spotify Web API. First, we get the seeds (a top artist and a top track of a user), by calculating the user’s expected preference to a particular track or artist according to his/her listening history. Due to ongoing interaction, the data is only updated once per day, thus two interfaces present the same seeds for the evaluation. Then, we take seeds as an input to call a recommendation service [3] that generates a play-list containing 20 songs matching similar artists and tracks. Each recommended song has a popularity score, genres, and audio features.

3.2 User interfaces

Figure 2 illustrates the design of the diversity-aware user interface which consists of two sections: section a) a visualization view shows an interactive visualization which allows users to explore songs by visualized attributes; section b) a list view shows all items in a list, and each of them is associated with a particular circle in visualization view. When the user clicks on a circle, the corresponding item in the list will be highlighted (red border) and vice versa. Each item in the list has a play icon and a thumb rating widget.

We hypothesize that visualizing additional meta-data of music such as audio features may result in higher diversity perceived by users with high MS and VM. Therefore, we designed the interfaces with two requirements. First, the visualizations should present multiple data dimensions effectively: in our case, we show two common attributes genres and popularity, and seven additional audio features. Second, the visualization should represent coverage by a particular attribute to reflect diversity, e.g., how the items are distributed by genres. Based on the above considerations, the bubble chart is selected as our primary visualization due to its good ability to present multidimensional data [23]. Moreover, to test our assumption, we also need to compare this relatively simple bubble chart (ComBub) with a baseline visualization. We consider a simple bubble chart (SimBub) as a good candidate since it meets the first requirement and uses almost the same visualization as ComBub. The visualizations were implemented with the D3.js library.

3.2.1 ComBub. Section a) of Figure 1 shows the design of ComBub that encodes the recommendations results in three ways. First, it uses a circle to represent each recommended song: the X and Y-axis are used to present two specified audio features. Second, the circle is color-coded for music genres, which allows users to distinguish song genres by their color. Third, the circle size (radius) is determined by the popularity score (from 1 to 100) which has been transformed by a visual square-root function. The function is defined as:

\[ R(p) = 6 \times \sqrt{\frac{p}{\pi}} \]  

(1)

where \( p \) is an item’s popularity score.

This encoding allows the user to inspect multiple dimension of the song simultaneously. The interface can be used to support advanced exploration for users such as popular pop songs with high danceability and high valence (happy, cheerful).
Most common interactions such as zooming and panning are supported. The details of a particular item will appear in a tooltip window when the mouse hovers over it. By clicking a circle, its associated item will be highlighted in the list synchronously. Below the plot, two drop-down menus are used to select audio features to visualize songs on the bubble chart. The scale of all audio features ranges from 0.0 to 1.0.

In summary, \textit{ComBub} allows users to specify two audio features to plot recommendations in two dimensions and inspect the details and distribution of genres and popularity as they wish. As explained above, the visualization is able to explain the diversity of recommendations from various aspects.

3.2.2 \textit{SimBub}. Figure 3 illustrates the design of \textit{SimBub}. To save space, the figure omits the recommendation list associated to the visualization that is identical to the one in section b) of Figure 2. We designed the simplest form of a bubble chart as a baseline for two reasons. First, this bubble chart represents items by labeled circles, which is a popular visualization among 13 common visualizations evaluated for visualizations at Internet scale [40]. Second, it can be seen as a variation of \textit{ComBub} without presenting audio features. Thus, it is easier for us to investigate the effects of the additional visualized audio features in \textit{ComBub}. Compared to \textit{ComBub}, this chart may be easier and sufficient for casual users to interpret and perceive diversity. In this sense, our study answers the question whether showing the additional audio features can lead to added value in terms of diversity and other investigated metrics of recommendations.

4 EVALUATION

To address our research questions, we conducted a user study to evaluate two visualizations in terms of diversity and other key metrics of recommender systems such as overall usability and user preferences.
4.1 Platform and Measurements

We chose Spotify as our experimental platform because it has become one of the largest on-line music service providers. Spotify provides access to more than 30 million songs and had more than 140 million monthly active users in June 2017. By leveraging the provided API, we generate recommendations according to real user data. In this study, we employ the native Spotify recommendation service to generate 20 songs based on a top artist and a top track of the user. Then, we implemented two visualizations on top of this recommender. In addition, the original list view interface and player widget were adopted to make playing music the same as in Spotify (See Figure 2).

We measure musical expertise by the Goldsmiths Musical Sophistication Index (Gold-MSI), consisting of ten questions on a 7-point Likert scale. The visual memory capacity is measured by the “Corsi block-tapping test” for short-term memory tasks. We also wrote a script to record user actions in a log file, which includes a timestamp, the times of checking song’s details (mouse hover), the times of exploring songs from the visualization (mouse click), and the rating of each song.

Moreover, to separate the effects of algorithm, we also control the actual diversity of recommendations to stay at a compared and moderate level. The actual diversity was measured by intra-list similarity (ILS) on music genres. We measure the similarity $C_o(b_k, b_e)$ between items $b_k, b_e$ based on the Jaccard similarity coefficient. Intra-list similarity for $a_i$’s list $P_{a_i}$ is defined as follows:

$$\text{ILS}(P_{a_i}) = \frac{\sum_{b_k \in \mathcal{P}_{a_i}} \sum_{b_e \in \mathcal{P}_{a_i} \cup b_k} C_o(b_k, b_e)}{2}$$

(2)

The Jaccard similarity is the number of common features for two sets A and B divided by the total number of features in the two sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

(3)

For all participants, recommendations shown in the two visualizations have a similar actual diversity calculated by ILS score (ComBub: Mean=22.46, SD=1.75, SimBub: Mean=22.07, SD=1.71). Lower scores obtained denote higher diversity.

4.2 Participants

We recruited 83 participants (Age: mean=28.3, SD = 6.7; Gender: 45.6% female) whose task approval rate is above 90% on Amazon MTurk. Four users were discarded because of self-contradictory responses found in reversal questions. All participant were compensated by 1.5 USD, and average completion time is 43 minutes. The majority of participants (69%) stated that they listened to on-line music every day. The remainder of the participants listened to on-line music at least once per week.

4.3 Evaluation Methods

We conducted a within-subjects study where participants evaluated two user interfaces (ComBub vs. SimBub). To minimize the learning effects, half of the participants evaluated two interfaces in a reverse order. The independent variable of the study is the type of visualization.

According to our research questions, there are three dependent variables in our study, perceived diversity, usability, and support to identify blind-spots. We modified the user-centric recommender evaluation questionnaire with thirteen statements (ResQue) to measure perceived diversity and interface usability. In particular, to measure categorical diversity, the following question was included: “S1: The items recommended to me are of various kinds.” The second question refers to item-to-item diversity by asking whether “S2: The items recommended to me are similar to each other.” Moreover, we measured novelty and serendipity by asking whether “S3: The recommender system helps me discover new songs.” and “S4: I haven’t heard of some songs in the list before.” respectively. Additionally, we also measured three other key aspects, usefulness (S5-S7), user’s attitude (S8-S10), and behavioral intention (S11-S13). The measured aspects by each statement are shown in Figure 4. All questions were measured by a 7-point Likert scale from “strongly disagree” (1) to “strongly agree” (7).

4.4 Hypotheses

Based on our research questions, we propose five hypotheses in our study.

H1: In general, ComBub supports the participants to gain higher perceived diversity than SimBub.

H2: ComBub supports the participants with high MS to gain higher perceived diversity than SimBub.

H3: ComBub supports the participants with high VM to gain higher perceived diversity than SimBub.

H4: ComBub is superior to SimBub in terms of overall usability.

H5: Both ComBub and SimBub are able to encourage the participants to identify the songs of under-explored genres.

4.5 Evaluation Procedure

First, we asked users to read a brief description of the study task and to watch a one minute video that shows all the functions and interactions supported by each visualization. After finishing the tutorial, the participants were required to sign in to our system with their Spotty accounts. Before showing the interface, they needed to fill out a pre-study questionnaire including demographic and background questions, and questions for testing musical sophistication and a link for measuring visual memory (Corsi test). To avoid cheating, participants needed to upload a screen-shot of the Corsi test score.

Participants were given the same task while testing the two visualizations: each participant needs to listen and rate all songs in the list with the possibility to explore recommendations through the interface. We required participants to spend at least ten minutes on testing each interface by disabling the questionnaire link. Despite the same algorithm and input seeds, the recommendations generated by Spotify vary between different requests. Thus, the potential influence of users’ familiarity with recommendation data is avoided. After using each visualization, the user was asked to fill out a post-study questionnaire. In the end, we asked all participants

---

3https://en.wikipedia.org/wiki/Spotify
4http://www.gold.ac.uk/music-mind-brain/gold-msi/
5http://www psytoolkit.org/experiment-library/corsi.html
We measured perceived diversity by aggregating the user ratings. We present the results according to each dependent variable.

The results do not show a significance difference between the two visualizations, ComBub (Mean=5.94, SD=0.46) and SimBub (Mean=5.26, SD=0.93), whereas participants with an average rating of 5 or more are categorized as high MS group (n=44). Table 1 shows diversity for two groups. In the group of high VM, the average score is 6.08 (SD=1.38). Similarly, average block span of 6.2 blocks (SD=1.3) in the Corsi test [22], shows the mean diversity for two groups. In the group of high VM, MS.

Table 1: Mean (SD) of diversity perceived by two groups of MS.

<table>
<thead>
<tr>
<th>MS levels</th>
<th>ComBub</th>
<th>SimBub</th>
</tr>
</thead>
<tbody>
<tr>
<td>High MS</td>
<td>5.94 (0.46)</td>
<td>5.26 (0.93)</td>
</tr>
<tr>
<td>Low MS</td>
<td>4.84 (0.67)</td>
<td>5.21 (0.78)</td>
</tr>
</tbody>
</table>

Table 2: Mean (SD) of diversity perceived by two groups of VM.

<table>
<thead>
<tr>
<th>VM levels</th>
<th>ComBub</th>
<th>SimBub</th>
</tr>
</thead>
<tbody>
<tr>
<td>High VM</td>
<td>5.91 (0.52)</td>
<td>5.08 (1.01)</td>
</tr>
<tr>
<td>Low VM</td>
<td>5.21 (0.80)</td>
<td>5.32 (0.78)</td>
</tr>
</tbody>
</table>

to indicate their preference for the two interfaces in terms of general preference, informativeness, usefulness, quality, and perceived diversity.

5 RESULTS

We present the results according to each dependent variable.

5.1 Perceived Diversity

We measured perceived diversity by aggregating the user ratings for four statements (S1-S4). Of note, we invert the rating of S2 because S2 is a reversal question. We performed a non-parametric Mann-Whitney test to analyze the significance of two visualizations on perceived diversity regardless of the variance of MS and VM. The results do not show a significance difference between the two visualizations, ComBub (Mean=5.45, SD=0.79) and SimBub (Mean=5.24, SD=0.86), on diversity (U = 2614.00, p = .88), thus H1 is not accepted. To verify hypotheses H2 and H3, we analyze the results of diversity by low and high levels of MS and VM separately.

5.1.1 Results Musical Sophistication. The average MS of all participants is 5.16 (SD=0.90). We quantified user’s MS based on their ratings of ten 7-point Likert scale questions. We categorized participants with an average rating of less than 5 as low MS group (n=35), whereas participants with an average rating of 5 or more are categorized as high MS group (n=44). Table 1 shows diversity for each visualization perceived by different user groups. In the group of high MS, a Mann-Whitney test shows that ComBub significantly outperforms SimBub (U = 500.50, p < .01), whereas for the group with low MS, SimBub is significantly better than ComBub (U = 429.00, p = .05). In addition, we performed a correlation analysis between perceived diversity and MS for the two visualizations. For ComBub, the results show a significantly positive correlation between MS and diversity (r=0.53, p<.01), and no significant correlation was found for SimBub (r=-0.009, p=.60). We can accept hypothesis H2.

5.1.2 Results Visual Memory Capacity. Healthy adults have an average block span of 6.2 blocks (SD=1.3) in the Corsi test [22], and in our study the average score is 6.08 (SD=1.38). Similarly, we categorized users into two groups, high VM (>6 blocks, n=27) and low VM (≤6 blocks, n=52) according to their scores. Table 2 shows the mean diversity for two groups. In the group of high VM, ComBub led to significantly higher perceived diversity than SimBub (U = 171.50, p < .01), while no significant result was found for the group of low VM (U = 1268.00, p = .58). The result of correlation analysis between VM and diversity shows a significantly positive correlation for ComBub (r=0.44, p<.01), and no significance was found for SimBub (r=0.007, p=.50). Hypothesis H3 can be accepted.

5.2 Overall Usability

We compare the overall usability of the two visualizations by performing Mann-Whitney tests for the responses to the post-study questionnaire. Figure 4 shows the average rating to each statement. In total, we found that for three statements, item-to-item diversity (S2), satisfaction (S8), and intention to reuse (S11), ComBub was rated significantly higher than SimBub. For the rest of the statements except facilitation (S5), despite no significance, ComBub still received higher ratings than SimBub. Thus, we can accept hypothesis H4.

Moreover, after evaluating the interfaces, participants were asked to show their preferences for the two visualizations in terms of five aspects. Figure 5 shows the distribution of preferences. In general, participants prefer ComBub (42%) over SimBub (29%), and 23% like both of them. Over half of the participants (54%) thought ComBub was more informative than SimBub. In terms of usefulness and quality, they showed an almost equal preference (35% vs 34%, 39% vs 34%) to ComBub and SimBub. 42% of users liked both visualizations in terms of perceived diversity.

5.3 Support to Identify Blind-spots

We recorded user actions with the two visualizations. When a user clicked a circle (song) we judged if the genre of this song had been visited. Then, we counted how many genres the participant had visited. The number of “hover” interactions indicates the counts of hovering the cursor to check the song information; and the number of “click” interactions is calculated by counting the number of clicks to highlight the associated item in the list. It is interesting to find that ComBub leads to significantly more user interactions (“hover”: U =
Figure 4: Results of the ResQue questionnaire. A cut-off value was set at 4.5 on the 7 point scale. The (*) sign means significant differences at the 5% level (p<.05) between visualizations. Of note, S2 is a reversal question.

Table 4: Mean number (SD) of each user action.

<table>
<thead>
<tr>
<th>index</th>
<th>ComBub</th>
<th>SimBub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of hover</td>
<td>58.28 (52.67)</td>
<td>8.18 (10.36)</td>
</tr>
<tr>
<td>Num. of click</td>
<td>6.51 (7.49)</td>
<td>3.85 (7.02)</td>
</tr>
<tr>
<td>Ratings</td>
<td>12.76 (3.82)</td>
<td>12.99 (3.96)</td>
</tr>
</tbody>
</table>

Visualizing the audio features of music has a limited impact on perceived diversity. The results suggest that these two visualizations are generally appropriate to support users in the perception of recommendation diversity. However, compared to SimBub, additional audio features of songs visualized in ComBub do not have significant added value for increasing perceived diversity if we disregard the effect of the individual traits MS and VM. Thus, we do not accept hypothesis H1 and answer the research question RQ1. We think that the understandability of what the features refer to could be a problem that hindered many people from profiting from the visualization of audio features. However, we do see a significant effect of audio features on item-to-item diversity (S2), which means ComBub was prone to help users to perceive the difference among items. We speculate that the visualization is good at helping users find how items are different in terms of their audio features.

Consider individual traits for designing diversity-aware recommendation interface. Overall, no significant difference was found between the two visualizations in terms of users’ perception of diversity. Instead, we find the most appropriate diversity-aware user interface depends on the two investigated individual traits: MS and VM.

We aimed to investigate the effects of two individual traits on user’s perception of diversity. We categorized the participants into two groups according to their MS and VM scores. The results indicate that for both users with high MS and users with high VM, their perceptions of diversity for recommendations in ComBub are significantly higher than the perceived diversity in SimBub, notwithstanding that the actual diversity of recommendations in ComBub is consistent with that in SimBub. Therefore, we could confirm the hypotheses H2, H3 and answer the research question RQ2 by finding the positive effects of MS and VM on perceived diversity. Of note, for those who have low MS, ComBub led to a significantly lower perception of diversity than SimBub. Thus, the additional audio features shown in ComBub do not seem to supply real benefits for users with low MS in terms of perception of diversity. Even the additional features in a visualization may result in higher cognitive load and a negative impact on user experience [4]. A similarly reversed result was not found for users with low VM.
The correlation analysis results show a significantly positive correlation between individual traits (MS, VM) and the perceived diversity in visualization ComBub; however, no such a significant correlation was found for SimBub. This result implies that an advanced visualization interface like ComBub allows experts to leverage their attribute knowledge to perceive higher diversity. In contrast, novices seem to prefer a simple interface that does not require intimate attribute knowledge such as the meaning of each audio features introduced in ComBub.

With this, we conclude that it is essential to consider the individual traits of users like MS and VM while presenting additional information like meta-data in a diversity-aware recommendation interface.

**Limited support in the exploration of blind-spots.** The subjective responses to the statements S3 and S4 (see Figure 4) show that participants were more likely to find songs which they had not yet listened to. Whereas the results of the user action log suggest that the visited songs through the visualizations only refer to a small portion (1/3) of all genres of recommendations. This may be explained by a relatively small number of clicks (see Table 4). It is interesting to find that the number of “hover” interactions is much higher than the number of “click” interactions. We surmise that most participants used the visualization as a tool of inspecting the details of items, or as a tooltip, rather than as an exploration tool. Despite the benefits of visualizations, users may still have checked the songs from the list directly, because interactions with visualization require extra click efforts. Thus, this addresses research question RQ3, we do not have evidence to support hypothesis H5.

7 LIMITATIONS

First, although we tried our best to minimize the potential harms to evaluation such as filtering workers and avoiding acquiescence bias by introducing contradictory statements, we cannot ignore the potential limitations [24] of using a crowd-sourcing platform like Amazon Mechanical Turk to evaluate a system with relatively complex tasks.

Second, the classification of genres used in our study contains 126 genres, which are defined by Spotify. However, we found some genres are quite similar, such as, “Punk” vs “Punk-Rock”. Therefore, the actual diversity should be lower than what we calculated.

Third, to ensure enough user engagement in testing two visualizations, we required users to spend at least ten minutes for each visualization and listen and rate all recommended songs. Thus, the recorded actions may not reflect the real user intention for clicking items on visualizations.

Despite these limitations, we are confident to argue that the music diversity perceived through user interfaces significantly depends on individual traits MS and VM, and considering these traits in the design of visualizations can lead to better support in enhancing diversity.

8 CONCLUSIONS AND FUTURE WORK

In addition to the approach that improves actual diversity by algorithms, increased attention has been paid to enhance user’s perception of diversity through user interfaces. A diversity-aware recommendation model (see Figure 1) suggests that the design of diversity-aware user interfaces should consider both user diversity and item diversity. Therefore, we hypothesize that users’ individual traits such as musical sophistication and visual memory may have a significant effect on the diversity perceived through user interfaces.

In particular, our study investigates such an effect with a music recommender user interface having two different visualizations for recommendations, ComBub and SimBub. Moreover, we are also interested in knowing if such a user interface is able to encourage users to explore blind-spots.

We presented an in-depth study to assess the value of showing additional item attributes (seven audio features) in the visualization, ComBub, for the purpose of increasing users’ perception of diversity. The results suggest that ComBub is particularly effective for enhancing the perceived diversity for users having high MS and high VM. In contrast, the baseline SimBub visualization (that shows fewer item attributes) is prone to increase the perception of diversity for the users having low MS. Furthermore, a strong positive correlation between individual traits and users’ perception of diversity was found in the ComBub visualization.

For our future work, we plan to validate our findings in other application domains. Moreover, we plan to improve the design of ComBub to see if it is able to encourage users to explore blind-spots in an exploration-oriented task.

9 ACKNOWLEDGEMENTS

Part of this research has been supported by the KU Leuven Research Council (grant agreement C24/16/017).

REFERENCES


