How Item Discovery Enabled by Diversity Leads to Increased Recommendation List Attractiveness

Bruce Ferwerda
Johannes Kepler University
bruce.ferwerda@jku.at

Mark P. Graus
TU Eindhoven
m.p.graus@tue.nl

Andreu Vall
Johannes Kepler University
andreu.vall@jku.at

Marko Tkalcic
Free University of Bolzano
marko.tkalcic@unibz.it

Markus Schedl
Johannes Kepler University
markus.schedl@jku.at

ABSTRACT

Applying diversity to a recommendation list has been shown to positively influence the user experience. A higher perceived diversity is argued to have a positive effect on the attractiveness of the recommendation list and a negative effect on the difficulty to make a choice. In a user study we presented 100 participants with several personalized lists of recommended music artists varying in levels of diversity. Participants were asked to assess these lists on perceived diversity and attractiveness, the experienced choice difficulty and discovery (i.e., the extent the list enriches their taste). We found that recommendation list attractiveness is influenced by two effects: 1) by diversity mediated through discovery; diverse recommendation lists are perceived to be more attractive if they enrich the user's taste or 2) by the list familiarity; a higher list familiarity contributes to a higher list attractiveness. We additionally revealed how individual differences (i.e., familiarity) moderate the effects found. Our results have implications on the composition of diversified recommendation lists. Specifically recommended items should contribute in extending and/or deepening the user's taste for the diversification to be effective.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI); User models; User studies;**

Keywords

Diversity; Recommender Systems; User-Centric Evaluation

1. INTRODUCTION & RELATED WORK

Recommender systems are usually designed in such a way that they provide the most relevant items to the user. However, this often results in a set of recommendations that are too similar to each other and thereby not cover the full spectrum of the user's interest [1]. Diversifying the recommendations can positively influence the user experience [5, 7, 8]. Recommendation diversity has been shown to contribute to the attractiveness of the recommendation list, which can reduce choice difficulty and increase choice satisfaction [5].

Diversification of recommendations has been shown to influence the user experience. Too similar recommendations or lack of diversity can have negative consequences [1]. Ziegler et al. [8] argued that recommendation list diversity can have positive effects on the user perception of the recommendation lists, and proposed a diversity algorithm to vary the recommendation list diversity without affecting the prediction accuracy. Willemsen et al. [7] investigated the influence of diversity on movie recommendations and found that diversity has a positive effect on the attractiveness of the recommendation set, choice difficulties, and eventually on the choice satisfaction. Although prior works have shown the positive effects of recommendation diversification, the reason behind this phenomenon has not yet been fully investigated. In this paper we take a deeper look at what leads to a higher attractiveness of the recommendation list, which subsequently leads to a lower choice difficulty. By testing music recommendation lists of varying diversity, we show that the attractiveness of the lists can be increased by list diversity *if* the items contribute to enriching the taste of the user. However, although the effect is weaker, list attractiveness can also be increased by presenting a list that users are familiar with. Our findings provide new insights on how to effectively create recommendation lists in order to increase list attractiveness and thereby decrease choice difficulties.

2. DATA PREPARATION & PROCEDURES

We created music recommendation lists of different diversification levels for participants in order to investigate the effects underlying the increase of recommendation list attractiveness. Since we created the recommendation lists off-line, we separated the study in two parts. In the first part participants were recruited and their *complete* Last.fm listening history was crawled in order to create the recommendation lists. After the lists were created, participants from the first part were invited for the second part where they were asked to assess the diversified recommendation lists.

We recruited 254 participants through Amazon Mechanical Turk for the first part of the study. Participation was re-

SAC 2017,April 03-07, 2017, Marrakech, Morocco

http://dx.doi.org/10.1145/3019612.3019899

stricted to those located in the United States with a very good reputation (≥95% HIT approval rate and ≥1000 HITs approved) and a Last.fm account with at least 25 listening events. A compensation of $1 was provided. We crawled the complete listening history of each participant and aggregated the listening events to represent artist and playcount (i.e., number of times listened to an artist).

To prepare the music recommendation lists for each participant, we complemented our data with the LFM-1b dataset. [1] This dataset consists of the complete listening histories of 120,322 Last.fm users from different countries. Since our participants were all located in the US, we only used US users of the LFM-1b dataset to complement our dataset with. This resulted in 10,255 additional users, which we aggregated into artist and playcount for each user. The final dataset consists of user, artist, and artist playcount triplets with 387,037 unique artists for the experiment.

We used the weighted matrix factorization algorithm by Hu et al. [4] on our final dataset to calculate the recommended items. This algorithm is specifically designed for datasets with implicit feedback (e.g., artist playcounts). We optimized the factorization hyper-parameters by conducting grid-search and picking the setting that yielded the best 5-fold cross-validated mean percentile rank. Specifically, using 20 factors, confidence scaling factor α=40, regularization weight λ=1000 and 10 iterations of alternating least squares, we achieved the best 5-fold cross-validated mean percentile rank of 1.78%. [2] Afterwards we factorized the whole user-artist triplets using this set of hyper-parameters.

The recommended items were diversified by using the topic diversification method of Ziegler et al. [8]. Using the latent features as the basis of diversification instead of additional metadata like genre information (as is done in content-based recommender systems) guarantees that diversity is manipulated in line with individual user preferences. Previous research demonstrated that this way of diversifying recommendations is perceived accordingly by users [7].

A greedy selection to optimize the intra-list similarity [1] was run on the top 200 recommended artists (i.e., the 200 artists with highest predicted relevance) to maximize the distances between item vectors in the matrix factorization space. This algorithm starts with a recommendation set consisting of the artist with highest predicted relevance. Iteratively, items are added to the recommendation set until it contains 10 items.

In each step of the iteration, for each candidate item i the sum of all distances from its item vector to each item vector in the recommendation set is calculated: $c_i = \sum_{j=1}^{z} d(i,j)$, where z is the number of items in the recommendation set and $d(i,j)$ is the Euclidean distance between two item vectors i and j. All candidate items are ranked based on decreasing value of c_i (P_{c_i}) and on predicted relevance (P_{r_i}). A weighting factor β following [8] is introduced to balance the trade-off between predicted relevance and diversity. For each candidate item the combined rank is calculated following $w_i^* = \beta * P_{c_i} + (1-\beta) * P_{r_i}$. The item with the highest combined rank is added to the recommendation set and the next step is taken until 10 items are selected.

[1]Available at http://www.cp.jku.at/datasets/LFM-1b/

[2]See [4] for details on the hyper-parameters and the definition of the mean percentile rank metric.

β was manipulated to achieve different levels of diversification. In the described implementation β=1 corresponds to maximum diversity, β=0 corresponds to maximum predicted relevance. We compared recommendation lists for different values of β in terms of the sum of distances between the latent features scores of items in the recommendation set and their average range. In terms of predicted rating, the list for β=0.4 showed to fall halfway between maximum relevance and maximum diversity. The final β levels for diversification were set at β=0 (low), β=0.4 (medium), and β=1 (high).

After the recommendation lists were created, emails were sent out to all participants to invite them for the second part of the study. We created a login screen so that we could retrieve the personalized recommendation lists for each participant. After the log in, the participant was sequentially presented with a recommendation list for three times, with each time a different level of diversity (i.e., low, medium, or high. The order of presentation was randomized). Each recommended artist was enriched with metadata from Last.fm (i.e., picture, genre, top 10 songs with the number of listeners and playcounts), which was shown when hovered over the name in the list. Additionally, example songs were provided by clicking on the artist name (new browser screen linked to the artist's YouTube page). Participants were asked to answer questions about perceived diversity, choice simplicity, recommendation attractiveness, music discovery, and list familiarity [3] before moving on to the next list. These questions needed to be answered for each of the three lists.

After the participant assessed all three recommendation lists, we performed a manipulation check by placing the three lists next to each other (randomly ordered) and asked the participant to rank order the lists by diversity. The study ended with a short questionnaire about their music expertise (Goldsmiths Musical Sophistication Index [6]) and their general preference strength (adapted from [7]).

There were 103 participants who returned for the second part of the study. We included several control questions to filter out careless contributions, which left us with 100 participants for the analyses. Age: 18-65 (median 28), gender: 54 male, 46 female, and were compensated with $2.

3. RESULTS

3.1 Manipulation Check

A Wilcoxon signed-rank test was used to test the perceived diversity levels of the recommendation lists. Results show an increase of perceived diversity by comparing the low diversity (M=1.28) against the medium (M=2.05, r=.60, Z=10.370, p<.001) and high condition (M=2.65, r=.80, Z=13.784, p<.001). A significant diversity increase was also found between medium and high (r=.45, Z=7.711, p<.001), indicating that the diversifications were correctly perceived.

3.2 Measures

Items in the questionnaire were assessed using a confirmatory factor analysis (CFA) to determine whether they convey

[3]Questions measuring perceived diversity, choice simplicity and recommendation attractiveness were adapted from [7]. Music discovery and list familiarity questions were created to understand the underlying effects of perceived diversity.

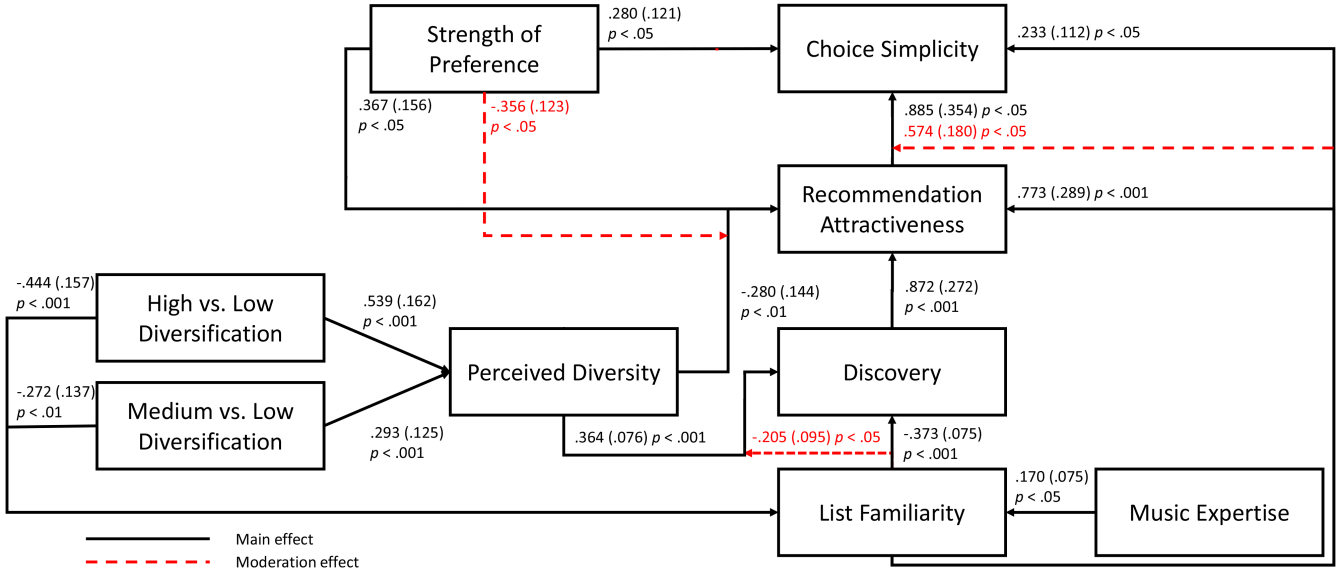

Figure 1: Path model. The numbers near the arrows denote the estimated means and the standard deviations.

the predicted constructs. After deleting questions with high cross-loadings and low commonalities, the model consisting of six constructs showed a good fit: $\chi^2(390)=695.4$, $p<.001$, $CFI=.96$, $TLI=.95$, $RMSEA=.05$. [4] The constructs with their items are shown below (5-point Likert scale; Disagree strongly-Agree strongly. Cronbach's alpha (α) and the average variance extracted (AVE) of each construct showed good values (i.e., $\alpha>.8$, $AVE>.5$), indicating convergent validity. The square root of the AVE for each construct is higher than any of the factor loading of the respective construct; indicating good discriminant validity. For the standardized music expertise questionnaire see [6].

Choice Simplicity ($AVE=.820$, $\alpha=.932$):

- I would find it easy to choose an artist to listen to because it stands out from the rest.
- I would find it difficult to choose an artist to listen to because all recommendations were equally bad.
- Many artists had comparable good aspects.

Recommendation Attractiveness ($AVE=.881$, $\alpha=.980$):

- I am satisfied with the list of recommended artists.
- In most ways the recommended artists were close to ideal.
- The list of artist recommendations meet my exact needs.
- I would give the recommended artists a high rating.
- The list of artists showed too many bad items.
- The list of artists was attractive.
- The list of recommendations matched my preferences.

Discovery ($AVE=.914$, $\alpha=.955$):

- The recommendations broadened my taste.
- The recommendations deepened my taste.

List Familiarity ($AVE=.871$, $\alpha=.953$):

- I am familiar with the recommended artists.
- I did not know the artists from the list.
- I already listen to the artists that were recommended.

Perceived Diversity ($AVE=.816$, $\alpha=.957$):

- The list of artists was varied.
- All the artists were similar to each other.
- Most artists were from the same genre.
- Many of the artists in the lists differed from each other.
- Artists differed a lot from each other on different aspects.

Strength of Preference (single item):

- I know what kind of music I like.

3.3 SEM Model

We created a path model with the subjective constructs of the CFA together with the music expertise construct using structural equation modeling (SEM) following the framework of Knijnenburg et al. [5] as a guideline (Figure 1). No order effects were observed in the order of presentation of the lists. The fit statistics show that the model has a good fit: $\chi^2(431)=701.4$, $p<.001$, $CFI=.96$, $TLI=.96$, $RMSEA=.04$. Only effects of $p<.05$ are included in the model. The medium and high diversity conditions are compared against the low condition in the results below to indicate the linear effect of the diversifications.

The medium and high diversification both show an increase in perceived diversity and an decrease in list familiarity compared to the low diversification. This confirms again that the diversification was correctly perceived.

A higher perceived diversity has a positive influence on the discovery of music (i.e., enriching one's taste). Discovery is furthermore negatively influenced by list familiarity; the discovery of new artists goes down when the familiarity goes up. The familiarity is in turn influenced by the music expertise of users; more expert users know more artists and thereby chances increase of knowing the ones in the list.

[4]Cutoff values are: $CFI>.96$, $TLI>.95$, $RSMEA<.05$ [3].

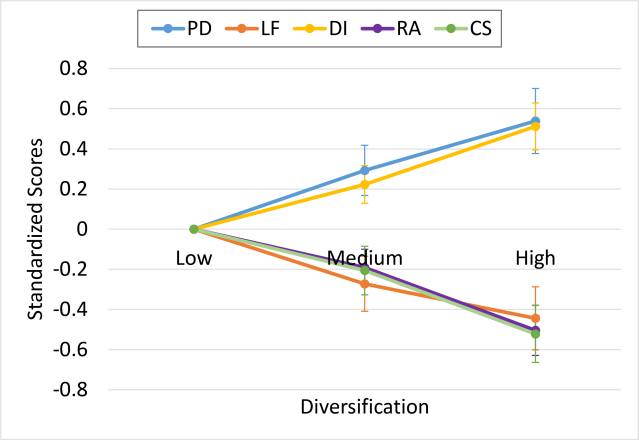

Figure 2: Marginal means with error bars of one std. error of the mean: perceived diversity (PD), list familiarity (LF), discovery (DI), recommendation attractiveness (RA), choice simplicity (CS).

The attractiveness of the recommendations is influenced by several constructs: perceived diversity, discovery, strength of preference, and list familiarity. Perceived diversity has a negative effect on the recommendation attractiveness, this effect gets stronger for those having a predefined preference. However, the effect becomes positive when it goes through the discovery of music. This implies that a higher degree of diversification has positive effects when the presented recommendations are able to enrich the music taste of users. Furthermore, strength of preference and list familiarity have both a positive effect on the perceived attractiveness.

The simplicity of making a choice is influenced by the recommendation attractiveness as well as by the strength of preference and list familiarity. The more attractive the recommendations are, the easier it becomes to make a choice. This effect is moderated by list familiarity. Preference strength and list familiarity plays a positive role on choice simplicity.

4. CONCLUSION & DISCUSSION

The general model we tested, as well as the results, show agreement with prior work on choice overload (e.g., [7]). However, the goal of this work was to get a deeper understanding of why a higher degree of diversity has positive effect on users' subjective evaluations of the recommendation list. We found that by increasing the list diversity, the perceived recommendation list attractiveness gets influenced. Our results indicate that recommendation list attractiveness is influenced by two effects: 1) by diversity mediated through discovery; diverse recommendation lists are perceived more attractive *if* they enrich the user's taste, or 2) by the list familiarity; a higher familiarity contributes to a higher attractiveness. Although these two ways were identified to increase list attractiveness, list diversification yields a stronger effect to increase the attractiveness and to subsequently increase choice simplicity. So it seems that for diversification to be effective, the list recommendations should contribute to enriching the users' taste by extending or deepening it.

Despite the new insights we provided on how diversification affects recommendation list attractiveness, our net effect of diversification on choice simplicity and recommendation attractiveness is negative (Figure 2). This in contrast to prior work [7] showing opposite effects. A possible explanation for this opposing effect may be a domain dependency (movies in cited work versus music in this study). Specifically the range in which we diversified (i.e., top 200) may be too broad for the music domain. Although we took the same range as prior work, music consists of more variable components (e.g., variations in genre, artist, context). By diversifying within the top 200, we may have created recommendation lists with items that are too far outside the spectrum of the user's taste; resulting in increasingly less attractive lists. We will further investigate the reason behind our negative net effect of diversification in future work.

Although prior work has shown that recommendation list diversity leads to a decrease in choice difficulties by increasing the list attractiveness [7], we provide new insights by showing that this effect occurs by enriching one's taste through the ability to discover new items. This should be taken into account when composing diversified recommendation lists.

We were unable to effectively measure choice satisfaction. Due to the within-subjects design of our experiment we did not ask participants to make a choice, but rather assess the recommendation lists. However, prior work (e.g., [7, 8]) has shown that choice simplicity leads to a higher choice satisfaction in recommender systems. We will address this question in future work. We will also explore additional moderating factors on discovery, such as personality traits [2].

5. ACKNOWLEDGMENTS

Supported by the Austrian Science Fund (FWF): P25655.

6. REFERENCES

[1] P. Castells, N. J. Hurley, and S. Vargas. Novelty and diversity in recommender systems. In *Recommender Systems Handbook*, pages 881–918. Springer, 2015.

[2] B. Ferwerda, M. Graus, A. Vall, M. Tkalcic, and M. Schedl. The influence of users' personality traits on satisfaction and attractiveness of diversified recommendation lists. In *EMPIRE*, page 43, 2016.

[3] L.-t. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

[4] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *ICDM*, 2008.

[5] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *UMUAI*, 2012.

[6] D. Müllensiefen, B. Gingras, L. Stewart, and J. Musil. Goldsmiths musical sophistication index, 2012.

[7] M. C. Willemsen, B. P. Knijnenburg, M. P. Graus, L. C. Velter-Bremmers, and K. Fu. Using latent features diversification to reduce choice difficulty in recommendation lists. *RecSys*, 11:14–20, 2011.

[8] C.-N. Ziegler, S. M. McNee, J. a. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. *WWW*, page 22, 2005.