QUANTITATIVE USER PERCEPTIONS OF MUSIC RECOMMENDATION LIST DIVERSITY

Kyle Robinson **Dan Brown**

David R. Cheriton School of Computer Science, University of Waterloo, Canada

kyle.robinson@uwaterloo.ca, dan.brown@uwaterloo.ca

ABSTRACT

Diversity is known to play an important role in recommender systems. However, its relationship to users and their satisfaction is not well understood, especially in the music domain. We present a user study: 92 participants were asked to evaluate personalized recommendation lists at varying levels of diversity. Recommendations were generated by two different collaborative filtering methods, and diversified in three different ways, one of which is a simple and novel method based on genre filtering. All diversified lists were recognised by users to be more diverse, and this diversification increased overall recommendation list satisfaction. Our simple filtering approach was also successful at tailoring diversity to some users. Within the collaborative filtering framework, however, we were not able to generate enough diversity to match all user preferences. Our results highlight the need to diversify in music recommendation lists, even when it comes at the cost of "accuracy".

1. INTRODUCTION

Music recommender systems play an ever-increasing role in individual listening habits as music consumption moves to online platforms and services such as Spotify, and Apple Music. Along with this growth has been an equivalent growth in research on how to better tailor music recommendations to match individual users' preferences and habits. Much of this research, especially in academia, depends on offline evaluations and metrics calculated on existing known listening histories as a proxy for real user satisfaction and list evaluation.

Along with metrics measuring the overall predictive ability (accuracy metrics) are offline evaluation metrics that measure additional qualities such as *Diversity*. These metrics are less standardised than accuracy metrics, and numerous definitions of each have been used in previous research [1, 2]. There is little public research on the effect of, and preference for, recommendation list diversity of actual music listeners.

We asked 92 online participants to evaluate three differently diversified personal recommendation lists from each of two different collaborative filtering recommendation algorithms. Participants were also asked questions about their preference for novel music and diversity as they relate to concepts discovered in the first study. We identified that accuracy and individual song ratings differed from overall *list satisfactio*n, and our implementation of *inner diversity* filtering resulted in higher levels of list satisfaction despite no significant decrease in perceived diversity. We also found that participants were less satisfied with the recommendations from a neural network model's recommendations despite its superior performance in offline testing accuracy. Finally, we found that none of our diversification methods resulted in too diverse recommendations, suggesting that were not able to match all users diversity preferences: some users wanted more diversity in their recommendations than we could provide.

2. BACKGROUND AND RELATED WORK

With novelty, coverage, and serendipity, diversity has long been identified as an important metric in providing satisfying automated recommendations to users [1]. Diversity has its origins in information retrieval, where it was used as a solution to ambiguous searches [3]. In recommender systems, diversity prevents over-personalization of recommendations to users in order to increase user satisfaction [1, 2]. Recommender system diversity has often been described as the opposite of *similarity* [4,5]. One definition of diversity in *music recommender systems* is intra-list diversity (ILD): the average pairwise dissimilarity of items by a similarity metric [5, 6]. There are many alternatives: modifications of ILD [7, 8] and novel approaches that do not rely on pairwise dissimilarity [9–11].

Vargas *et al.* use the distributions of genres in a user listening histories and recommendations to satisfy three *properties* of diversity: genre coverage, redundancy, and size-awareness [10]. Oliveira *et al.* similarly seek to *Pareto*-optimize a set of self-selected *aspects* of diversity: contemporaneity, locality, gender, and genre [12]. None of these methods are evaluated with any users, though they outperform other methods on the defined metrics.

Anderson *et al.* found that use of personalized recommendations leads to a reduction in overall consumption diversity, diversity was related to higher user retention, and users' consumption diversity was increased by a migration away from personalized recommendations [13]. Holtz *et al.* found similar results with podcast recommendations [14]. Finally, Hansen *et al.* examine different methods of shifting consumption on one large music platform towards more diversity [15]. Although these works provide vital information on the current diversity of users on music platforms, they do so using commercial metrics such as retention and consumption. We provide a more foundational view of diversity for user satisfaction and perception.

3. METHODOLOGY

We trained our own recommendation models to control for all aspects of recommendation and diversification.

3.1 Recommendation Overview

3.1.1 Data

We extended the publicly available LastFM data set created for our previous work by retroactively topping up each user's listening history [16]. For each LastFM username we collected up to 10,000 new song Listening Events (LEs) starting from July 2020 and working back to their latest LE in the existing data set. Users who did not have any new LEs during this period were removed from the data set. Tracks in this data set are identified using unique artist and track name tuples. The total un-processed and updated data set consists of 520,134,112 unique LEs, and 15,804,356 unique artist-track tuples recorded by 50,440 users over a period of roughly 2 years during 2019 and 2020.

To remove noise, we eliminated tracks which were listened to 10 or less times. This filtering resulted in a drastic reduction in unique artist-track tuples to 2,817,819 (82.2% decrease), while only modestly reducing the number of LEs to 488,528,514 (6% decrease) and the number of users to 50,437 ($< 0.01\%$ decrease).

In the filtered data set, the median number of LEs per user is 9,857, the 25$^{\text{th}}$ percentile is 4,663, and the 75th$^{\text{th}}$ percentile is 14,277. The user-track-interaction matrix contains 176,151,310 non-zero entries (play counts) across 2,817,819 unique tracks, resulting in a 50,437×2,817,819-sparse matrix. Entries in this matrix correspond to the number of unique times a user (row) played the track (column). An anonymized version of this updated data is available upon request.

Data was split into training, validation, and test sets using weak generalization, where user-item interactions are sampled at random from the entire dataset to form the subsets. This differs from the strong generalization used by Liang *et al.* which samples entire users resulting in each user occuring in only one data subset [17]. We used weak generalization because matrix factorization can not efficiently deal with a large number of unseen users. Data was split into train, validation, and test subsets by successively splitting by a ratio of 85/15.

3.1.2 Algorithms

We chose two collaborative filtering recommendation algorithms designed for implicit feedback data sets: Alternating Least Squares matrix factorization (ALS) [18], and Variational Autoencoders for Collaborative filtering (MultVAE) [17]. The results of both algorithms are presented as a form of replication for diversity, and we plan to contrast the overall performance of both models in another work. We provide a brief overview of how these algorithms work in practice, and refer readers to the original papers for detailed descriptions and mathematical processes.

ALS uses classical matrix factorization, and has been used frequently in recommendation research and production [18–20]. The algorithm generates recommendations by factorizing a large sparse matrix of user and item playcounts to compute a low-rank matrix approximation. The factorizations give a vector for each user and item. The product of any user vector and item vector represents relevance.Vectors for unseen users can be calculated using their listening history and the latent item factorizations. This method is known as the *fold-in* method [21].

In addition to generating recommendations, the column factorizations give latent features representing each song.

MultVAE is a modern neural network approach based on a Variational Autoencoder architecture. It is the only neural network approach identified by Dacrema *et al.* to outperform basic top-*n* recommendation algorithms using various measurements of accuracy on commonly used benchmark data sets [22]. MultVAE passes a dense input vector with length equal to the total number of recommendable items (x) through an encoder (g_ϕ) to a lower-dimensional latent representation (z), and then through a decoder (f_θ) which has an inverse architecture to the encoder. The general architecture of MultVAE is:

$$x \longrightarrow g_\phi \longrightarrow z \longrightarrow f_\theta \longrightarrow x'$$

The authors suggest that g_ϕ and f_θ consist of 0 or 1 densely connected perceptron layers with a dimensionality of 600, and the dimensionality of z to be 200. The dimensionality of x and x' is equal to the number of items in the database. The vector x' gives expected play counts which we can sort in decreasing order and select top-*n* items from.

3.1.3 Hyperparameter Optimization and Training

For the general performance analysis, hyperparameter optimization, and baseline comparison we adopt binary Normalized Discounted Cumulative Gain (NDCG) [4]. Discounted Cumulative Gain (DCG) is based on recall and defined as: $DCG = \sum_{i=1}^{k} \frac{rel_i}{\log(i+1)}$ where rel is a binary value representing whether the recommendation at rank i appears in the unseen portion of the users listening history. The denominator then *discounts* the relevance based on how far from rank 1 it appears. NDCG for one user is: $NDCG = \frac{DCG}{DCG'}$ where DCG' is the ideal DCG: rel_i is always equal to 1. Total NDCG@k is the average value across all users for some defined list length k.

We optimized ALS hyperparameters using randomized search over 60 iterations. The best performance on valida-

Model	Validation NDCG@100	Test NDCG@100
ALS	0.217	0.325
MultVAE	0.223	0.349

Table 1: Final results of both recommendation algorithms. Note that the test results reflect models trained using the combined training and validation data sets.

tion data was achieved using 224 factors, $\lambda = 1$, $\alpha = 1$, after 98 iterations.

Our implementation of MultVAE was based on the original author's Tensorflow 1 implementation, and a PyTorch implementation by James Le [1]. Due to the large number of unique tracks in our data set, full cross-validation of MultVAE architectures was not computationally feasible. We instead trained a number of different models and architectures concurrently based on the original authors results. The best performance on validation data was achieved using 0 hidden encoder/decoder layers, annealing cap of 1, 10000 annealing steps, learning rate of 0.001, and batch size of 500 over 250 epochs. We implemented early stopping based on NDCG@100, but it was not triggered. The final dimensionality of the model was:

$$[2,817,819] \longrightarrow [200] \longrightarrow [2,817,819]$$

Both models were retrained on the combined training and validation data, and evaluated on the unseen test data. The final evaluation results can be seen in Table 1.

ALS, generates a new latent user vector using their listening history and the existing latent item vectors. We multiply this new user vector with all item vectors to generate item relevance. For MultVAE we feed the user's listening history through the trained network and obtain a new vector containing each item's relevance. The relevance values from each list are then sorted in decreasing order to form top-*n* lists.

3.2 Item Features

We calculate diversity with latent item features generated from ALS matrix factorization. To lessen the effect of popularity on latent features, each track's feature vector was ℓ_2-normalized to unit-length. Item distances were computed using simple Euclidean distance.

3.3 Music Recommendation Lists

We used three different techniques to generate top-10 music recommendation lists for both recommendation algorithms, giving 6 different top-10 recommendation lists per user. Recommendation lists generated using ALS are prefixed with *als*, and recommendation lists generated using MultVAE are prefixed with *vae*.

3.3.1 Control (als, vae)

Our control recommendation lists consist of the raw ranked output from each recommendation algorithm after removing tracks which appeared in the user's listening history. This gives the metrics reported in Table 1.

3.3.2 Maximally Diverse (als_max_div, vae_max_div)

We generated these recommendation lists using the greedy ILD diversification method described by Ziegler *et al.* using $\beta = 1$ [6]. This greedy diversification algorithm starts with the maximally diverse track from some larger recommendation list; we start with the top-1000 recommendations from each model. The algorithm incrementally adds the track maximally distant from the already selected tracks until the list is of the desired length. This method ensures that the tracks are not only maximally distant from all other tracks, but that the final recommendation list traverses multiple extremes in the item feature space. We do not consider the relevance ranking within the top-1000 recommendations when generating diverse recommendation lists; this corresponds to setting $\beta = 1$ in the original diversification process.

3.3.3 Filtered Diverse (als_filt_div, vae_filt_div)

We also use filtered diverse lists, where the top-1000 recommendations are filtered based on the user's existing listening history. This aims to better align recommendations with user preference for *inner diversity* identified in existing research [16]. We considered two methods for filtering recommendations: feature clustering, and genre filtering.

For feature clustering, we tried to remove tracks too distant from existing LEs. We clustered user LE history into *n* groups, and filtered recommendations which fall outside the clustering. This approach proved unsuccessful.

For genre filtering we remove recommendations in genres which do not appear in the user's existing listening history. We used Spotify artist genre tags, and defined a track's genres as the genres of that track's artist retrieved from Spotify.

For each user, we identified all genres which appear in the user's listening history and their frequencies. Next, we find the most diverse track among the top-1000 recommendations (the first track in the Maximally Diverse list) and its genres. The user's genre list is searched for this track's genres, and we save the lowest frequency found (or 0 if none) to be the user's genre threshold. We remove from the top-1000 recommendation list any tracks with a single genre either not in the user's hash table, or with a frequency below the found threshold. We run greedy diversification on the filtered list.

3.4 User Study

Our interactive user study consisted of a pre-interaction survey on personal music consumption, discovery, and preference, followed by 6 personalized top-10 music recommendation lists as described in Section 3.3. The recommendation lists included a 5-point Likert evaluation for each track, and questions on the recommendation list as

[1] The original authors' code can be found at https://github.com/dawenl/vae cf. Permission to use Le's code was obtained through email correspondence; his implementation can be found at https://github.com/khanhnamle1994/MetaRec

a whole using the same 5-point Likert scale. The music recommendations were displayed as 30 second song previews using Spotify Play Button widgets.[2] The study was hosted as an online web-app which collected participant LastFM data and generated recommendations while participants completed the surveys.

We completed a pilot study with participants recruited at our institution in order to test the system before completing the primary study. The pilot study included a post-interview on their experience with the system. No significant concerns were discovered during the pilot study. Primary study participants were recruited through Amazon Mechanical Turk, whose terms of service prohibit asking workers (participants) to register for a service, or log into an existing service. We therefore required workers to have a LastFM account in order to participate, and specified such in the HIT description, the HIT layout, and as a question on the consent form. After obtaining informed consent, we had workers provide their LastFM username which we used to obtain their public listening history. We also required that the LastFM account had at least 50 LEs recorded in the last 6 months. To verify ownership without requiring a login, workers were given 3 attempts to name one artist they had listened to in the previous 6 months. Pilot participants were compensated $10CAD, and primary participants were compensated $4USD.

Participants were shown recommendation lists using a balanced Latin square design to control for differences in recommendation list order.

4. RESULTS

4.1 Data and Demographics

We recruited 9 pilot participants, and 97 primary participants. Only primary participant data was used for analysis.

Five participants were removed for completing lists too quickly, resulting in a final participant count of 92. The proceeding results include only these 92 participants. Completion times for *vae* and *als* were observably lower than for diversified lists.

The median participant age was 29, the youngest was 19 and the oldest 62; 50 identified as male (54%), 40 identified as female (43%), and 2 identified as non-binary (2%).

Only tracks in our base training data set can be used to generate recommendations and be recommended. The median count of LEs per participant was 857 before removing tracks not in the base data set, and 627 after. This is compared to a median value of 3110 for users in the base data set.

4.2 Pre-Interaction Survey

In addition to demography, the pre-interaction survey asked questions focused on music consumption, discovery, and music recommendation preferences.

Participants agreed that their diversity preference depended on who makes the recommendations, the quality of the recommendations, what they are doing, and their mood. Almost 50% of participants disagreed or strongly disagreed that their location was important to how they felt at a given time about music diversity.

We also asked yes/no questions about diversity preference–most participants selected "yes" for all with the notable exception of the question: "Do you want music recommendations outside of genres you like?", for which 32% indicated they were unsure, and 15% responded “no”. This question also serves as a parallel to the ideas of *inner* and *outer* diversity preference.

4.3 Recommendation List Evaluation

We assign labels to the recommendation list evaluation questions based on the order in which they were presented to participants. These questions, their labels, and their responses can be seen in Figure 1.

4.4 List Comparisons

We preformed a Friedman test on the distributions of responses between each list for all LQ and found that at least one list type's distribution differed significantly for each LQ ($p < 0.001$ for all). A post-hoc Nemenyi test is performed to identify which list's distributions differ from each other. The results of the post-hoc tests are visualised in Figure 1 as black bars connecting significantly different distributions. Note that statistical significance is found more readily among LQ0 because there are 10 samples per list, and we used Dunn's tests instead of Nemenyi test[3].

In the responses to rating questions (LQ0) and satisfaction questions (LQ3, LQ4), the control *als* recommendations were consistently rated more positively than the *vae* recommendations (LQ0: $p \leq 0.001$, LQ4: $p = 0.005$). The *als_max_div* recommendations were also consistently rated higher than *vae_max_div* (LQ0: $p \leq 0.001$, LQ3: $p = 0.007$, LQ4: $p \leq 0.001$). Regarding the filtered lists, *als_filt_div* and *vae_filt_div* were rated similarly or better than their un-diversified counterparts in list satisfaction (LQ3, LQ4) despite receiving less positive individual track ratings (LQ0). Filtered lists also performed similar to or better than their maximally diverse counterparts in all cases.

We also examined the distributions of responses to LQ0 and LQ3 by list type using a Kruskal Wallis test, and found significant differences among the control lists ($p \leq 0.001$), which highlights a clear distinction between track ratings and overall list satisfaction, especially for control lists.

The diversity results (LQ1, LQ2, LQ5, LQ6, LQ7) show that all diversified lists were recognised to be significantly more diverse, and to portray a wider range of genres than than their non-diversifed controls ($LQ1, LQ5, LQ7 : p \leq 0.001$). No significant differences in perceived diversity or genre range were found between filtered diverse and maximally diverse lists. Participants did not find any list

[2] https://developer.spotify.com/documentation/widgets/generate/play-button/

[3] We preform Kruskal-Wallis and Dunn's (with Bonferroni adjustment) tests for LQ0 instead of Friedman and Nemenyi due to the unbalanced data. While this test is typically used for independent samples, we are unaware of a better non-parametric alternative.

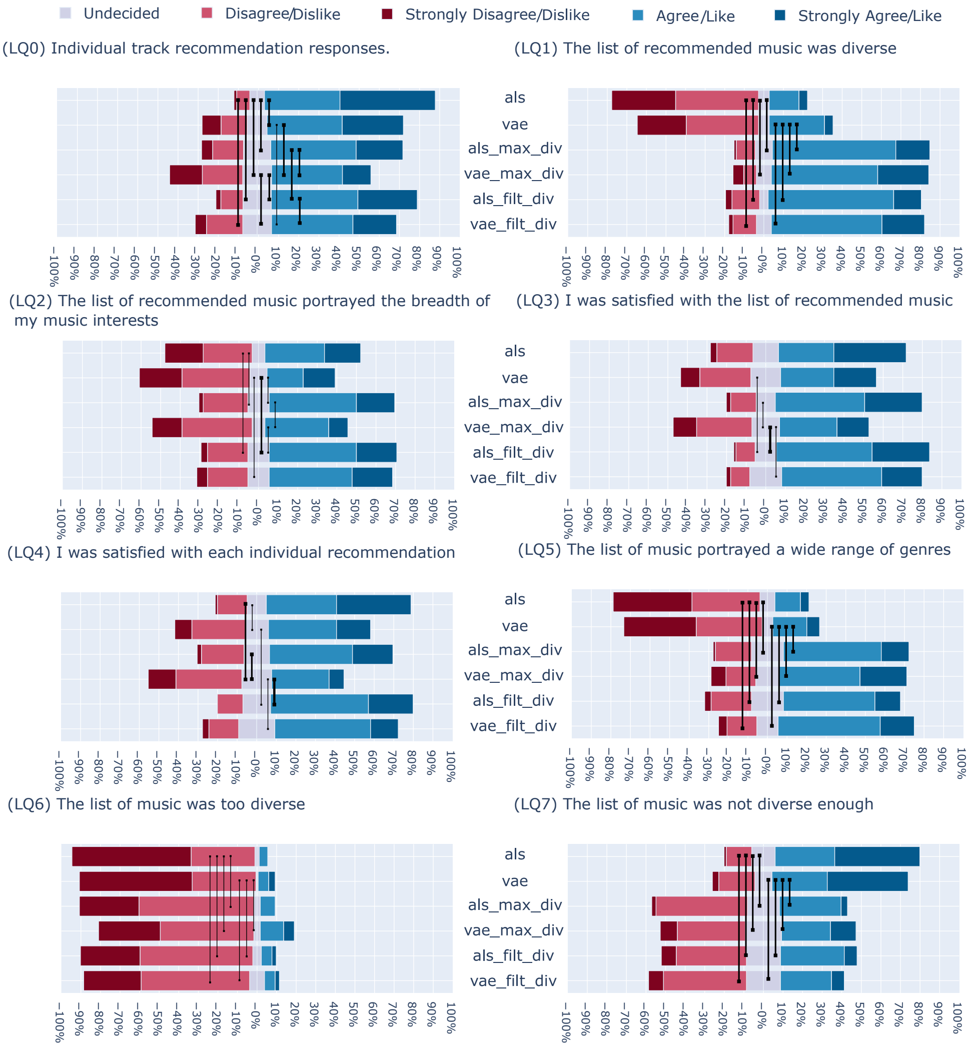

Figure 1: Responses to Likert questions on recommendation lists. Black bars connect significantly different distributions (thick: $p \leq 0.001$, thin:$p \leq 0.05$). LQ0 ratings used Like/Dislike, while all others used Agree/Disagree.

to be overly diverse, though they did feel more strongly that the control lists were not overly diverse as compared to most diversified lists ($LQ6 : p \leq 0.05$). The filtered *als_filt_div* and *vae_filt_div* lists most consistently portrayed the breadth of participants' music interests (LQ2).

An additional Kruskal-Wallis test was performed on the distributions of responses to LQ1 and LQ5 for all list types with no statistically significant results, supporting the idea that one way users perceive diversity is genre range.

4.5 Summary of Statistically Significant Results

We found statistically significant differences among recommendation algorithms and list generation approaches. In general, recommendations from the ALS model were more satisfying than those from the VAE model, and filtered lists performed similar to or better than control, and maximally diversified lists from the same model. List satisfaction and individual track ratings also differed significantly.

In analysing diversity responses we found that all diverse recommendation lists were recognised as such, and filtered lists were found to be just as diverse as maximally diversified ones. Filtered lists also most consistently conveyed the breadth of participants interests. Additionally, participant responses on genre range mirrored their evaluations of diversity, and no list types were found to be too diverse.

In the next section, we explore what our results suggest about how to build good music recommender systems.

5. DISCUSSION

5.1 Satisfaction

In recommender systems research, the quality of a recommendation list is often inferred from some accuracy measures computed on known data sets [4]. Our results show that accuracy does **not** tell the whole story. MultVAE outperformed ALS on the base data set using NDCG@100 (Table 1), but participant responses on individual recommendations showed markedly higher satisfaction from the ALS model with an ostensibly lower test accuracy. This gap in satisfaction only grows larger for the maximally diverse recommendations generated from top-1000 lists. We plan to explore this dichotomy through additional and more exhaustive offline evaluation in another manuscript.

5.2 Diversity

Previous research has identified that mood and context are important factors in determining how much diversity a user wants [16], and this is further supported by user responses on diversity preference which showed that users identify mood and context (among other factors) as important. The difference in responses between 'what I am doing' vs. 'where I am' emphasise the difference between context and location. A user may be working out if they are detected at a gym, but the location alone is not as significant as the action.

Previous work on optimising diversity levels in recommendation lists has also depended heavily on accuracy measurements [1, 2, 4]. Our results suggest that the difference between control and diversified lists is not well portrayed in individual recommendation ratings. Although maximally diverse lists did result in lower individual track ratings (LQ0), there was no detectable impact on overall list satisfaction (LQ3). Despite strong statistical evidence that both filtered and maximally diversified lists were significantly more diverse. It is especially important to keep in mind that the maximally diverse lists were created using a beta value of 1 from all top-1000 participant recommendations. Either the additional diversity of the lists made up for a decrease in the quality of each recommendation, or the top-1000 recommendations are all of a relatively high quality.

5.3 Genre and Filtering

The nearly identical responses to questions about list diversity and the range of genres further solidify the relationship between the two concepts [10, 16]. When users are asked to evaluate the diversity of a music recommendation list, genre is clearly one of the primary factors they consider.

Overall, the filtered recommendation lists performed as well or better than the maximally diversified lists for satisfaction while also portraying similar levels of diversity. When maximally diverse recommendations were good (as for *als_max_div*) the filtering had no statistically significant impact on list satisfaction or diversity. Alternatively, when maximally diverse recommendations were poor (as for *vae_max_div*) the filtering had a sizeable positive impact on satisfaction without impacting perceived diversity.

The filtered and maximally diverse lists can be viewed as simple implementations of a system for inner and outer diversity.

5.4 Diversity and Personalization

The ILD method we chose is arguably the simplest such diversity metric. Despite its simplicity, our results add to the existing evidence that increasing ILD is perceived by users as increasing diversity, this time in the domain of music [7, 20]. In fact, the significant negative impact of this diversification method was only observed in the Mult-VAE recommendations, and was removed through genre filtering.

We extend this one step further by noting that the filtered recommendations were generated with at $\beta = 1$. Filtering can result in positive satisfaction even with maximal ILD, suggesting that any and all values of β will present viable recommendation lists for each user. This may simplify the task of selecting an optimal level of diversification based on mood and context.

5.5 Pushing Diversity Further

We were unable to generate recommendation lists which reached outside the bounds of our participant's diversity preferences. Even maximally diverse recommendations were not seen as too diverse. It is very hard to generate overly diverse recommendations using either model. In an ideal collaborative filtering system, diversity preference would be implicitly considered. Also, some users prefer *outer diversity*: recommendations which differ from their existing listening preferences.

Since hyper-parameter optimization of recommendation models make use of accuracy measurements such as NDCG@k (Section 3.1.3) which incentivize only recommendations in training users' hidden listening history; most existing recommender systems, because they so strongly focus on accuracy, are unlikely to make risky recommendations.

In order to generate **truly** diverse music recommendations that match user preference, we first need to understand the extents of their preferences for diversity. The idea of recommending surprising items is typically associated with the related beyond-accuracy metric of serendipity [23]. It is easy to equate user preference for *outer diversity* to a preference for serendipity, but this does not explain why even maximally diverse recommendations were not too diverse. Perhaps by extending beyond the top-1000 most relevant recommendations, we can find more diverse recommendations.

If collaborative filtering algorithms do not generate adequate levels of diversity, then are they really working towards generating better music recommendations for users?

5.6 Summary

Our results highlight the large disconnect between offline and online accuracy and diversity evaluations of music recommender systems. Through a sizeable within-subjects study, we evaluated two collaborative filtering algorithms and found that the offline accuracy–and even the user provided track ratings–were not good indicators of overall list satisfaction.

Diversity continues to be an important topic of discussion in recommender systems. Our genre filter-based diversification approach enabled satisfying and diverse recommendations within users' existing preferences despite using a simple diversity definition. We found success in modeling diversity based on user ideas of the term, and then asking them to evaluate it. In doing so, we brought to light the limited diversity contained within collaborative filtering recommendation algorithms.

6. REFERENCES

[1] M. Kaminskas and D. Bridge, "Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 1, pp. 1–42, 2016.

[2] M. Kunaver and T. Požrl, "Diversity in recommender systems – A survey," *Knowledge-Based Systems*, vol. 123, pp. 154–162, 2017.

[3] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 659–666.

[4] A. Gunawardana and G. Shani, "Evaluating recommender systems," in *Recommender Systems Handbook, Second Edition*. Springer, 2015, pp. 265–308.

[5] K. Bradley and B. Smyth, "Improving recommendation diversity," in *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, 2001, pp. 85–94.

[6] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the World Wide Web Conference, WWW*, 2005, p. 22.

[7] S. Castagnos, A. Brun, and A. Boyer, "When Diversity Is Needed... But Not Expected!" in *International Conference on Advances in Information Mining and Management*, 2013, pp. 44–50.

[8] A. L'Huillier, S. Castagnos, and A. Boyer, "Understanding usages by modeling diversity over time," in *CEUR Workshop Proceedings*, vol. 1181, 2014, pp. 81–86.

[9] S. Vargas, "New Approaches to Diversity and Novelty in Recommender Systems," in *FDIA'11: Proceedings of the Fourth BCS-IRSG conference on Future Directions in Information Access*, 2011, pp. 8–13.

[10] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells, "Coverage, redundancy and size-awareness in genre diversity for recommender systems," in *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*, 2014, pp. 209–216.

[11] D. M. Fleder and K. Hosanagar, "Recommender systems and their impact on sales diversity," in *EC'07 - Proceedings of the Eighth Annual Conference on Electronic Commerce*, 2007, pp. 192–199.

[12] R. S. Oliveira, C. Nóbrega, L. B. Marinho, and N. Andrade, "A multiobjective music recommendation approach for aspect-based diversification." in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 2017, pp. 414–420.

[13] A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas, "Algorithmic Effects on the Diversity of Consumption on Spotify," in *Proceedings of the World Wide Web Conference, WWW*, apr 2020, pp. 2155–2165.

[14] D. Holtz, B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral, "The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify," in *EC 2020 - Proceedings of the 21st ACM Conference on Economics and Computation*, jul 2020, pp. 75–76.

[15] C. Hansen, R. Mehrotra, C. Hansen, B. Brost, L. Maystre, and M. Lalmas, "Shifting Consumption towards Diverse Content on Music Streaming Platforms," in *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, aug 2021, pp. 238–246.

[16] K. Robinson, D. G. Brown, and M. Schedl, "User insights on diversity in music recommendation lists," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020, pp. 446–453.

[17] D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara, "Variational autoencoders for collaborative filtering," in *Proceedings of the World Wide Web Conference, WWW*, 2018, pp. 689–698.

[18] Y. Hu, C. Volinsky, and Y. Koren, "Collaborative filtering for implicit feedback datasets," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 263–272, 2008.

[19] B. Frederickson, "Fast Python Collaborative Filtering for Implicit Datasets." https://github.com/benfred/implicit, 2019.

[20] B. Ferwerda, M. Graus, A. Vall, M. Tkalčič, and M. Schedl, "The Influence of Users' Personality Traits on Satisfaction and Attractiveness of Diversified Recommendation Lists," in *Proceedings of the 4th Workshop on Emotions and Personality in Personalized Services (EMPIRE 2016)*, 2016, pp. 43–47.

[21] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Incremental singular value decomposition algorithms for highly scalable recommender systems," in *Fifth International Conference on Computer and Information Science*, vol. 1, no. 012002, 2002, pp. 27–8.

[22] M. F. Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? A worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 101–109.

[23] P. Castells, N. J. Hurley, and S. Vargas, "Novelty and diversity in recommender systems," in *Recommender Systems Handbook, Second Edition*. Springer, 2015, pp. 881–918.