IMPACT OF LISTENING BEHAVIOR ON MUSIC RECOMMENDATION

Katayoun Farrahi
Goldsmiths, University of London
London, UK
k.farrahi@gold.ac.uk

Markus Schedl, Andreu Vall, David Hauger, Marko Tkalčič
Johannes Kepler University
Linz, Austria
firstname.lastname@jku.at

ABSTRACT

The next generation of music recommendation systems will be increasingly intelligent and likely take into account user behavior for more personalized recommendations. In this work we consider user behavior when making recommendations with features extracted from a user's history of listening events. We investigate the impact of listener's behavior by considering features such as play counts, "mainstreaminess", and diversity in music taste on the performance of various music recommendation approaches. The underlying dataset has been collected by crawling social media (specifically Twitter) for listening events. Each user's listening behavior is characterized into a three dimensional feature space consisting of play count, "mainstreaminess" (i.e. the degree to which the observed user listens to currently popular artists), and diversity (i.e. the diversity of genres the observed user listens to). Drawing subsets of the 28,000 users in our dataset, according to these three dimensions, we evaluate whether these dimensions influence figures of merit of various music recommendation approaches, in particular, collaborative filtering (CF) and CF enhanced by cultural information such as users located in the same city or country.

1. INTRODUCTION

Early attempts in collaborative filtering (CF) recommender systems for music content have generally treated all users as equivalent in the algorithm [1]. The predicted score (i.e. the likelihood that the observed user would like the observed music piece) was a weighted average of the K nearest neighbors in a given similarity space [8]. The only way the users were treated differently was the weight, which reflected the similarity between users. However, users' behavior in the consumption of music (and other multimedia material in general) has more dimensions than just ratings. Recently, there has been an increase of research in music consumption behavior and recommender systems that draw inspiration from psychology research on personality. Personality accounts for the individual difference in users in their behavioral styles [9]. Studies showed that personality affects rating behavior [6], music genre preferences [11] and taste diversity both in music [11] and other domains (e.g. movies in [2]).

The aforementioned work inspired us to investigate how user features intuitively derived from personality traits affect the performance of a CF recommender system in the music domain. We chose three user features that are arguably proxies of various personality traits for user clustering and fine-tuning of the CF recommender system. The chosen features are *play counts*, *mainstreaminess* and *diversity*. Play count is a measure of how often the observed user engages in music listening (intuitively related to extraversion). Mainstreaminess is a measure that describes to what degree the observed user prefers currently popular songs or artists over non-popular (and is intuitively related to openness and agreeableness). The diversity feature is a measure of how diverse the observed user's spectrum of listened music is (intuitively related to openness).

In this paper, we consider the music listening behavior of a set of 28,000 users, obtained by crawling and analyzing microblogs. By characterizing users across a three dimensional space of play count, mainstreaminess, and diversity, we group users and evaluate various recommendation algorithms across these behavioral features. The goal is to determine whether or not the evaluated behavioral features influence the recommendation algorithms, and if so which directions are most promising. Overall, we find that recommending with collaborative filtering enhanced by continent and country information generally performs best. We also find that recommendations for users with large play counts, higher diversity and mainstreaminess values are better.

2. RELATED WORK

The presented work stands at the crossroads of personality-inspired user features and recommender systems based on collaborative filtering.

Among various models of personality, the Five-factor model (FFM) is the most widely used and is composed of the following traits: *openness*, *conscientiousness*, *extraversion*, *agreeableness* and *neuroticism* [9]. The personality theory inspired several works in the field of recommender systems. For example, Pu et al. [6] showed that user rating behavior is correlated with personality factors. Tkalčič et al. [13] used FFM factors to calculate similarities in a CF recommender system for images. A study by

Rentfrow et al. [11] showed that scoring high on certain personality traits is correlated with genre preferences and other listening preferences like diversity. Chen et al. [2] argue that people who score high in openness to new experiences prefer more diverse recommendations than people who score low. The last two studies explore the relations between personality and diversity. In fact, the study of diversity in recommending items has become popular after the publishing of two popular books, *The Long Tail* [4] and *The Filter Bubble* [10]. However, most of the work was focused on the trade-off between recommending diverse and similar items (e.g. in [7]). In our work, we treat diversity not as a way of presenting music items but as a user feature, which is a novel way of addressing the usage of diversity in recommender systems.

The presented work builds on collaborative filtering (CF) techniques that are well established in the recommender systems domain [1]. CF methods have been improved using context information when available [3]. Recently, [12] incorporated geospatial context to improve music recommendations on a dataset gathered through microblog crawling [5]. In the presented work, we advance this work by including personality-inspired user features.

3. USER BEHAVIOR MODELING

3.1 Dataset

We use the "Million Musical Tweets Dataset" [1] (MMTD) dataset of music listening activities inferred from microblogs. This dataset is freely available [5], and contains approximately 1,100,000 listening events of 215,000 users listening to a total of 134,000 unique songs by 25,000 artists, collected from Twitter. The data was acquired crawling Twitter and identifying music listening events in tweets, using several databases and rule-based filters. Among others, the dataset contains information on location for each post, which enables location-aware analyses and recommendations. Location is provided both as GPS coordinates and semantic identifiers, including continent, country, state, county, and city.

The MMTD contains a large number of users with only a few listening events. These users are not suitable for reliable recommendation and evaluation. Therefore, we consider a subset of users who had at least five listening events over different artists. This subset consists of 28,000 users.

Basic statistics of the data used in all experiments are given in Table 1. The second column shows the total amount of the entities in the corresponding first row, whereas the right-most six columns show principal statistics based on the number of tweets.

3.2 Behavioral Features

Each user is defined by a set of three behavioral features: play count, diversity, and mainstreaminess, defined next. These features are used to group users and to determine how they influence the recommendation process.

[1] http://www.cp.jku.at/datasets/MMTD

Play count The play count of a user, $P(u)$, is a measure of the quantity of listening events for a user u. It is computed as the total number of listening events recorded over all time for a given user.

Diversity The diversity of a user, $D(u)$, can be thought of as a measure which captures the range of listening tastes by the user. It is computed as the total number of unique genres associated with all of the artists listened to by a given user. Genre information was obtained by gathering the top tags from Last.fm for each artist in the collection. We then identified genres within these tags by matching the tags to a selection of 20 genres indicated by Allmusic.com.

Mainstreaminess The mainstreaminess $M(u)$ is a measure of how mainstream a user u is in terms of her/his listening behavior. It reflects the share of most popular artists within all the artists user u has listened to. Users that listen mostly to artists that are popular in a given time window tend to have high $M(u)$, while users who listen more to artists that are rarely among the most popular ones tend to score low.

For each time window $i \in \{1 \dots I\}$ within the dataset (where I is the number of all time windows in the dataset) we calculated the set of the most popular artists A_i. We calculated the most popular artists in an observed time period as follows. For the given period we sorted the artists by the aggregate of the listening events they received in a decreasing order. Then, the top k artists, that cover at least 50% of all the listening events of the observed period are regarded as popular artists. For each user u in a given time window i we counted the number of play counts of popular artists $P_i^p(u)$ and normalized it with all the play counts of that user in the observed time window $P_i^a(u)$. The final value $M(u)$ was aggregated by averaging the partial values for each time window:

$$M(u) = \frac{1}{I} \sum_{i=1}^{I} \frac{P_i^p(u)}{P_i^a(u)} \tag{1}$$

In our experiments, we investigated time windows of six months and twelve months.

Table 3 shows the correlation between individual user features. No significant correlation was found, except for the mainstreaminess using an interval of six months and an interval of twelve months, which is expected.

3.3 User Groups

Each user is characterized by a three dimensional feature vector consisting of $M(u)$, $D(u)$, $P(u)$. The distribution of users across these features are illustrated in Figures 1 and 2. In Figure 3, mainstreaminess is considered with a 6 month interval. The results illustrate the even distribution of users across these features. Therefore, for grouping users, we consider each feature individually and divide users between groups considering a threshold.

For mainstreaminess, we consider the histogram of $M(u)$ (Figure 2 for a 6 month (top) and 12 month (bottom)) in making the groups. We consider 2 different cases for grouping users. First, we divide the users into 2 groups according to the median value (referred to as M6(12)-median-G1(2)).

Level	**Amount**	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Users	27,778	5	7	10	27.69	17	89,320
Artists	21,397	1	1	2	35.95	9	11,850
Tracks	108,676	1	1	1	7.08	4	2,753
Continents	7	9	4,506	101,400	109,900.00	142,200	374,300
Countries	166	1	12	71	4,633.00	555	151,600
States	872	1	7	40	882.00	195	148,900
Counties	3557	1	2	10	216.20	41	191,900
Cities	15123	1	1	5	50.86	16	148,900

Table 1. Basic dataset characteristics, where "Amount" is the number of items, and the statistics correspond to the values of the data.

	RB	C_{cnt}	C_{cry}	C_{sta}	C_{cty}	C_{cit}	CF	CC_{cnt}	CC_{cry}	CC_{sta}	CC_{cty}	CC_{cit}
P-top10	10.28	**11.75**	11.1	5.70	5.70	5.70	11.22	10.74	10.47	5.89	5.89	5.89
P-mid5k	1.33	1.75	2.25	2.43	1.46	1.96	4.47	**4.59**	4.51	3.56	1.96	2.56
P-bottom22k	0.64	0.92	1.10	1.03	0.77	1.07	1.85	**1.95**	**1.95**	1.56	0.96	1.16
P-G1	0.45	0.67	0.72	0.68	0.44	0.56	1.13	**1.26**	1.17	0.78	0.26	0.35
P-G2	0.65	1.32	1.34	1.01	0.69	0.92	1.71	**1.78**	1.77	1.32	0.80	0.89
P-G3	1.08	2.04	2.02	1.88	1.30	1.73	3.51	**3.60**	3.59	2.90	1.68	2.16
D-G1	0.64	0.85	1.16	1.04	0.87	0.88	2.22	**2.24**	2.16	1.59	0.97	0.93
D-G2	0.73	0.93	1.05	1.23	0.84	1.02	2.04	**2.21**	2.20	1.68	0.98	1.08
D-G3	0.93	1.63	1.49	1.56	0.93	1.41	2.49	2.56	**2.59**	2.03	1.08	1.54
M6-03-G1	0.50	0.88	0.95	0.96	0.64	0.88	1.76	**1.84**	**1.84**	1.43	0.81	1.00
M6-03-G2	1.34	2.73	2.43	2.22	1.49	2.00	3.36	**3.50**	**3.50**	2.81	1.67	2.08
M6-median-G1	0.35	0.58	0.62	0.65	0.48	0.61	1.35	**1.46**	1.45	1.04	0.56	0.66
M6-median-G2	1.25	2.49	2.89	2.25	1.47	1.97	3.14	3.27	**3.29**	2.67	1.66	2.07
M12-05-G1	1.35	2.02	2.27	2.25	1.50	1.93	2.90	3.02	**3.04**	2.47	1.54	1.94
M12-05-G2	0.36	0.59	0.69	0.61	0.41	0.57	1.30	**1.38**	**1.38**	1.01	0.52	0.66
M12-median-G1	0.36	0.62	0.71	0.64	0.43	0.59	1.41	**1.50**	**1.50**	1.10	0.56	0.71
M12-median-G2	1.34	2.09	2.33	2.34	1.57	2.01	3.10	3.24	**3.26**	2.66	1.67	2.10

Table 2. Maximum F-score for all combinations of methods and user sets. C refers to the CULT approaches, CC to CF_CULT; cnt indicates continent, cry country, sta state, cty county, and cit city. The best performing recommenders for a given group are in bold.

Second, we divide users into 2 groups for which borders are defined by a mainstreaminess of 0.3 and 0.5, respectively, for the 6 month case and the 12 month case (referred to as M6(12)-03(05)-G1(2)). These values were chosen by considering the histograms in Figure 2 and choosing values which naturally grouped users. For the diversity, we create 3 groups according to the 0.33 and 0.67 percentiles (referred to as D-G1(2,3)). For play counts, we consider 2 different groupings. The first is the same as for diversity, i.e. dividing groups according to the 0.33 and 0.67 percentiles (referred to as P-G1(2,3)). The second splits the users according to the accumulative play counts into the following groups, each of which accounts for approximately a third of all play counts: top 10 users, mid 5,000 users, bottom 22,000 users (referred to as P-top10(mid5k,bottom22k)).

4. RECOMMENDATION MODELS

In the considered music recommendation models, each user $u \in U$ is represented by a list of artists listened to $A(u)$. All approaches determine for a given seed user u a number K of most similar neighbors $V_K(u)$, and recommend the artists listened to by these $V_K(u)$, excluding the artists $A(u)$ already known by u. The recommended artists $R(u)$ for user u are computed as $R(u) = \bigcup_{v \in V_K(u)} A(v) \setminus A(u)$ and $V_K(u) = \text{argmax}^K_{v \in U \setminus \{u\}} sim(u, v)$, where argmax^K_v denotes the K users v with highest similarities to u. In considering geographical information for user-context models, we investigate the following approaches, which differ in the way this similarity term $sim(u, v)$ is computed. The following approaches were investigated:

	D(u)	**M(u) (6 mo.)**	**P(u)**
D(u)	-	0.119	0.292
M(u) (12 mo.)	0.069	0.837	0.013
P(u)	0.292	0.021	-

Table 3. Feature correlations. Note due to the symmetry of these featuers, mainstreaminess is presented for 6 months on one dimension and 12 months on another. Overall, none of the features are highly correlated other than the mainstreaminess 6 and 12 month features, which is expected.

CULT: In the cultural approach, we select the neighbors for the seed user only according to a geographical similarity computed by means of the Jaccard index on listening distributions over semantic locations. We consider as such

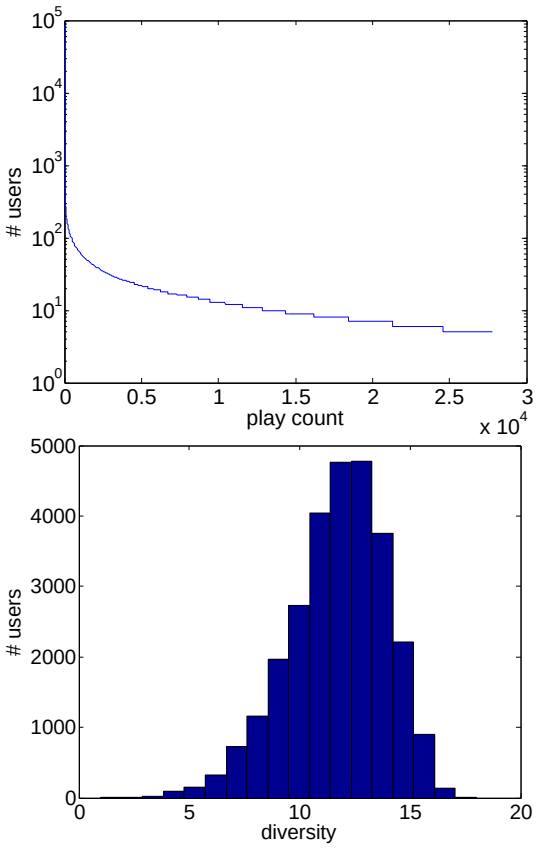

Figure 1. Histogram of (top) play counts (note the log scale on the y-axis) and (bottom) diversity over users.

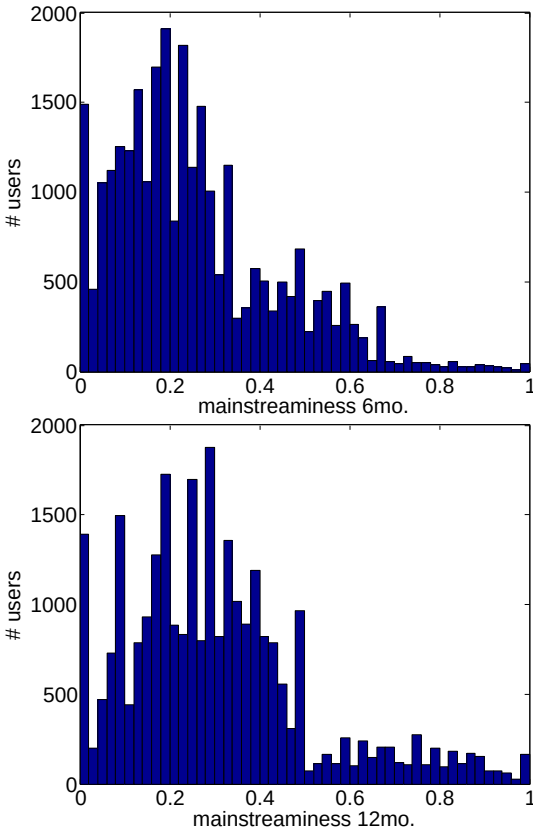

Figure 2. Histogram of mainstreaminess considering a time interval of (top) 6 months and (bottom) 12 months.

semantic categories continent, country, state, county, and city. For each user, we obtain the relevant locations by computing the relative frequencies of his listening events over all locations. To exclude the aforementioned geoentities that are unlikely to contribute to the user's cultural circle, we retain only locations at which the user has listened to music with a frequency above his own average [2]. On the corresponding listening vectors over locations of two users u and v, we compute the Jaccard index to obtain $sim(u, v)$. Depending on the location category user similarities are computed on, we distinguish CULT_continent, CULT_country, CULT_state, CULT_county, and CULT_city.

CF: We also consider a user-based collaborative filtering approach. Given the artist play counts of seed user u as a vector $\vec{P}(u)$ over all artists in the corpus, we first omit the artists that occur in the test set (i.e. we set to 0 the play count values for artists we want our algorithm to predict). We then normalize $\vec{P}(u)$ so that its Euclidean norm equals 1 and compute similarities $sim(u, v)$ as the inner product between $\vec{P}(u)$ and $\vec{P}(v)$.

CF_CULT: This approach works by combining the CF similarity matrix with the CULT similarity matrix via pointwise multiplication, in order to incorporate both music preference and cultural information.

RB: For comparison, we implemented a random baseline model that randomly picks K users and recommends the artists they listened to. The similarity function can thus be considered $sim(u, v) = rand\,[0,1]$.

[2] This way we exclude, for instance, locations where the user might have spent only a few days during vacation.

5. EVALUATION

5.1 Experimental Setup

For experiments, we perform 10-fold cross validation on the user level. For each user, we predict 10% of the artists based on the remaining 90% used for training. We compute precision, recall, and F-measure by averaging the results over all folds per user and all users in the dataset. To compare the performance between approaches, we use a parameter N for the number of recommended artists, and adapt dynamically the number of neighbors K to be considered for the seed user u. This is necessary since we do not know how many artists should be predicted for a given user (this number varies over users and approaches). To determine a suited value of K for a given recommendation approach and a given N, we start the approach with $K = 1$ and iteratively increase K until the number of recommended artists equals or exceeds N. In the latter case, we sort the returned artists according to their overall popularity among the K neighbors and recommend the top N.

5.2 Results

Table 2 depicts the maximum F-score (over all values of N) for each combination of user set and method. We decided to report the maximum F-scores, because recall and

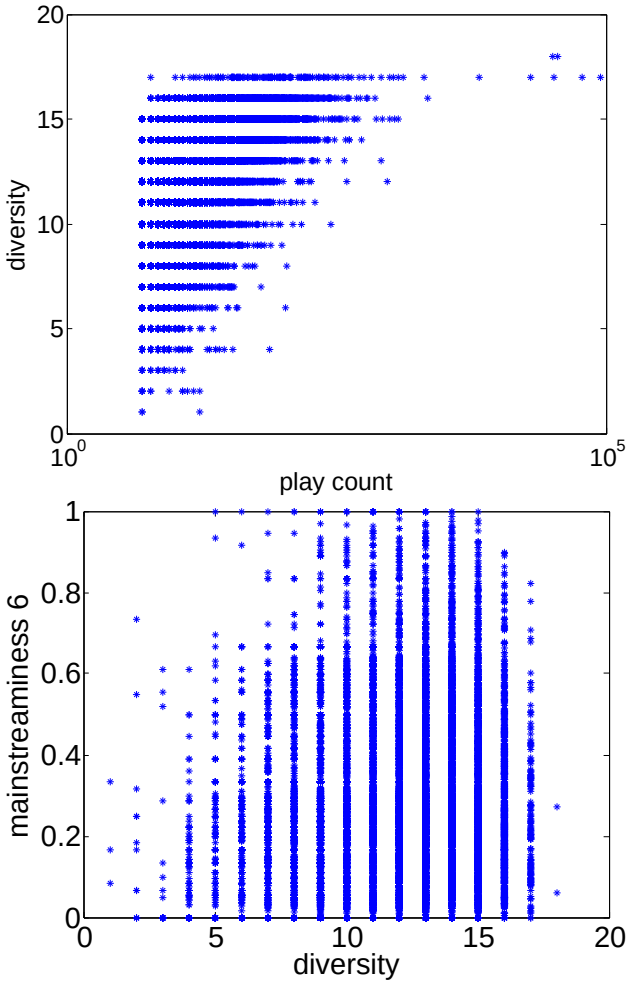

Figure 3. Users plot as a function of (top) $D(u)$ vs $P(u)$ and (bottom) $M(u)$ (6 months) vs $D(u)$. Note the log scale for $P(u)$ only. These figures illustrate the widespread, even distribution of users across the feature space.

precision show an inverse characteristics over N. Since the F-score equals the harmonic mean of precision and recall, it is less influenced by variations of N, nevertheless aggregate performance in a meaningful way. We further plot precision/recall-curves for several cases reported in Table 2. In Figure 4, we present the results of all of the recommendation algorithms for one group on the play counts. For this case, the CF approach with integrated continent and country information performed best, followed by the CF approach. Predominantly, these three methods outperformed all of the other approaches for the various groups, which is also apparent in Table 2. The only exception was the P-top10 case, where the CULT_continent approach outperformed CF approaches. However, considering the small number of users in this subset (10), the difference of one percentage point between CULT_continent and CF_CULT_continent is not significant. We observe the CF approach with the addition of the continent and country information are very good recommenders in general for the data we are using.

Now we are interested to know how the recommendations performed across user groups and respective features.

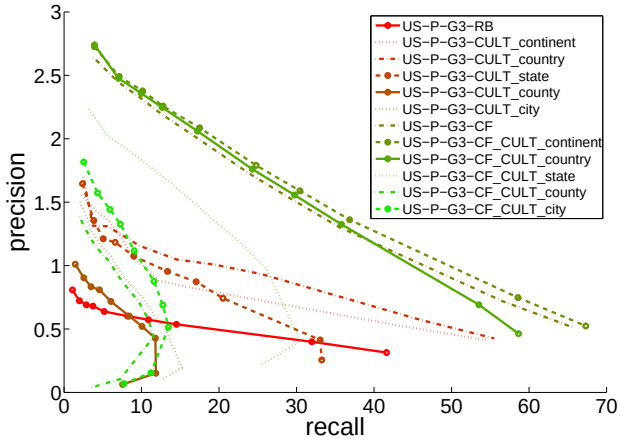

Figure 4. Recommendation performance of investigated methods on user group P-G3.

In terms of play counts, we observe as the user has a larger number of events in the dataset, the performance increases significantly (P-G3 and P-top10). This can be explained by the fact that more comprehensive user models can be created for users about whom we know more, which in turn yields better recommendations.

Also in terms of diversity, there are performance differences across groups given a particular recommender algorithm. Especially between the high diversity listeners D-G3 and low diversity listeners D-G1, results differ substantially. This can be explained by the fact that it is easier to find a considerable amount of like-minded users for seeds who have a diverse music taste, in technical terms, less sparse $A(u)$ vector.

When considering mainstreaminess, taking either a 6 month or 12 month interval does not appear to have a significant impact on recommendation performance. There are minor differences depending on the recommendation algorithm. However, in general, the groups with larger mainstreaminess (M6-03-G2, M6-med-G2, M12-med-G2) always performed much better for all approaches than the groups with smaller mainstreaminess. It hence seems easier to satisfy users with a mainstream music taste than users with diverging taste.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we consider the role of user listening behavior related to the history of listening events in order to evaluate how this may effect music recommendation, particularly considering the direction of personalization. We investigate three user characteristics, play count, mainstreaminess, and diversity, and form groups of users along these dimensions. We evaluate several different recommendation algorithms, particularly collaborative filtering (CF), and CF augmented by location information. We find the CF and CF approaches augmented by continent and country information about the listener to outperform the other methods. We also find recommendation algorithms for users with large play counts, higher diversity, and higher

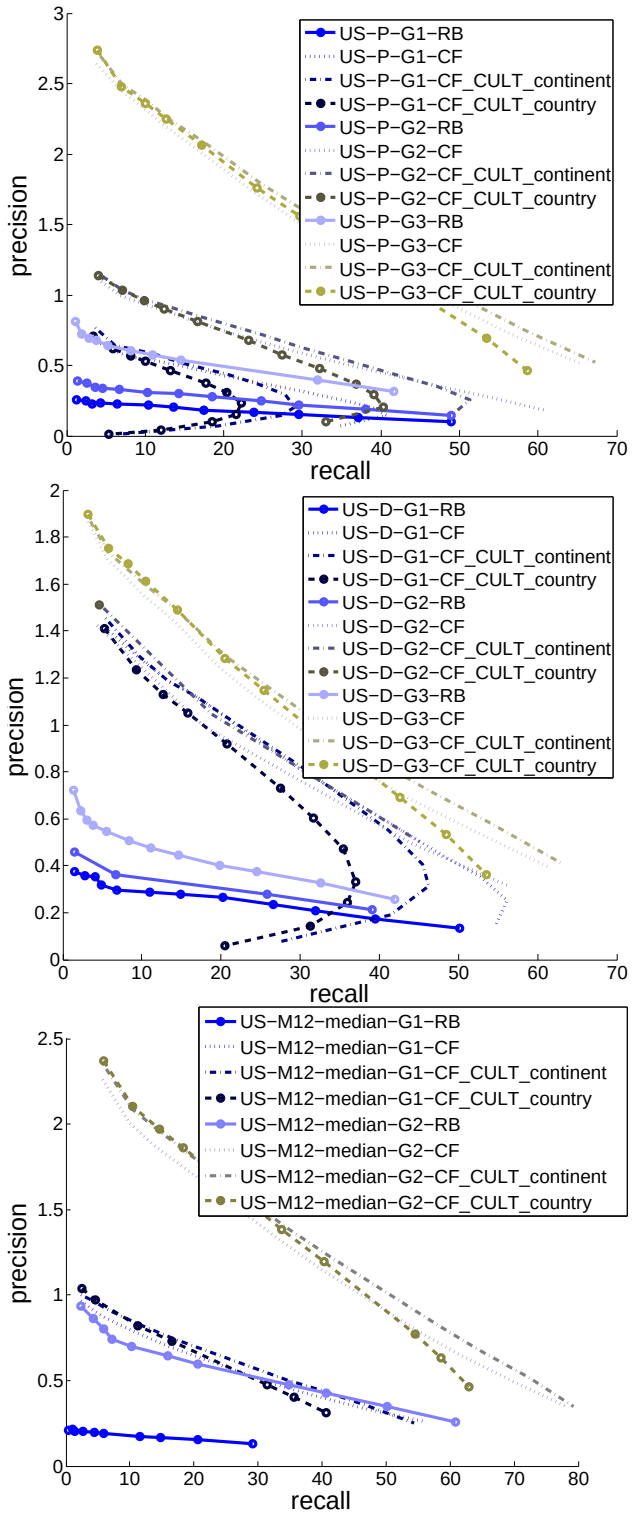

Figure 5. Precision vs. recall for play count (top), diversity (middle), and mainstreaminess with a 12 month interval (bottom) experiments over groups and various recommendation approaches.

mainstreaminess have better performance.

As part of future work, we will investigate content-based music recommendation models as well as combinations of content-based, CF-based, and location-based models. Additional characteristics of the user, such as age, gender, or musical education, will be addressed, too.

7. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Funds (FWF): P22856 and P25655, and by the EU FP7 project no. 601166 (“PHENICX”).

8. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] L. Chen, W. Wu, and L. He. How personality influences users’ needs for recommendation diversity? *CHI ’13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA ’13*, 2013.

[3] N. Hariri, B. Mobasher, and R. Burke. Context-aware music recommendation based on latent topic sequential patterns. In *Proc. ACM RecSys ’12*, New York, NY, USA, 2012.

[4] M. Hart. The long tail: Why the future of business is selling less of more by chris anderson. *Journal of Product Innovation Management*, 24(3):274–276, 2007.

[5] D. Hauger, M. Schedl, A. Košir, and M. Tkalčič. The Million Musical Tweets Dataset: What Can We Learn From Microblogs. In *Proc. ISMIR*, Curitiba, Brazil, November 2013.

[6] R. Hu and P. Pu. Exploring Relations between Personality and User Rating Behaviors. *1st Workshop on Emotions and Personality in Personalized Services (EMPIRE)*, June 2013.

[7] N. Hurley and M. Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Trans. Internet Technol.*, 10(4):14:1–14:30, March 2011.

[8] J. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, March 2012.

[9] R. McCrae and O. John. An Introduction to the Five-Factor Model and its Applications. *Journal of Personality*, 60(2):175–215, 1992.

[10] E. Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[11] P. Rentfrow and S. Gosling. The do re mi’s of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.

[12] M. Schedl and D. Schnitzer. Hybrid Retrieval Approaches to Geospatial Music Recommendation. In *Proc. ACM SIGIR*, Dublin, Ireland, July–August 2013.

[13] M. Tkalčič, M. Kunaver, A. Košir, and J. Tasič. Addressing the new user problem with a personality based user similarity measure. *Joint Proc. DEMRA and UMMS*, 2011.