Tailoring Music Recommendations to Users by Considering Diversity, Mainstreaminess, and Novelty

Markus Schedl
Department of Computational Perception
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

David Hauger
Department of Computational Perception
Johannes Kepler University
Linz, Austria
david.hauger@jku.at

ABSTRACT

A shortcoming of current approaches for music recommendation is that they consider user-specific characteristics only on a very simple level, typically as some kind of interaction between users and items when employing collaborative filtering. To alleviate this issue, we propose several user features that model aspects of the user's music listening behavior: diversity, mainstreaminess, and novelty of the user's music taste. To validate the proposed features, we conduct a comprehensive evaluation of a variety of music recommendation approaches (stand-alone and hybrids) on a collection of almost 200 million listening events gathered from *Last.fm*. We report first results and highlight cases where our diversity, mainstreaminess, and novelty features can be beneficially integrated into music recommender systems.

Categories and Subject Descriptors

Information systems [**Information retrieval**]: Music recommendation

Keywords

Music Information Retrieval, Music Recommendation, Recommender Systems, User Modeling, Evaluation

1. MOTIVATION AND RELATED WORK

Music recommendation has become a popular research and business topic over the past few years [3, 10]. While the importance of incorporating user characteristics and contextual aspects into recommender systems has been acknowledged many times, among others in [12, 14, 1], work that looks into this matter in the domain of music recommendation is still scarce.

Baltrunas et al. [2] focus on music recommendation in cars, taking into account traffic, road, driver, and weather conditions. Cheng and Shen [4] propose a music recommender that exploits location, listening history, music descriptors, and music popularity trends. Wang et al. [13] propose a mobile music recommender that predicts user activity from time, acceleration, and ambient noise, and selects music from a set of tracks that have been labeled in advance by the user to fit a certain activity. A similar approach is taken by Park et al. [8] who learn emotional states from weather, light level, and temporal data, and again recommend music by matching with pre-labeled tracks. These works concentrate on factors other than the previous music listening behavior when modeling the user, ignoring the potential usefulness of decent features that describe aspects of the listeners' music consumption.[1]

We already investigated in [11] a variety of user demographics to filter results of different music recommendation algorithms. We found significant improvements when tailoring music recommendations to users of similar age, living in the same country (in particular, for the US and Russia), and enjoying either folk, blues, or jazz music. In the work at hand, we develop more elaborate listening-specific user attributes and center our recommendation models around these. We further investigate a content-based recommender and a variety of hybrids and provide detailed results.

In the remainder of the paper we first present the employed music recommendation approaches (Section 2). We then introduce the proposed user characteristics (Section 3), before detailing the used dataset and experimental setup, and discussing the results (Section 4). Conclusions and possible extensions round off this paper (Section 5).

2. RECOMMENDATION APPROACHES

We investigate stand-alone approaches as well as hybrids, which are detailed in the following. Each user u has a listening profile L_u which contains all items listened to — artists in our case. Furthermore, u is assigned a normalized playcount vector $\vec{p}_u$ containing the number of listening events over all items I in the corpus. This vector is normalized so that its Euclidean norm equals 1. For the content-based approaches, each item i is assigned a TF·IDF vector $\vec{w}_i$ which we construct by (i) gathering all *Last.fm* tags and tag weights of i, (ii) treating the tags of i as a document and the corresponding tag weights as term frequencies, (iii) filtering tags that occur for less than 100 items (i.e. $< 0.01\%$ of all

[1]We do not consider as decent features the result of feeding the listening histories of users into a collaborative filtering approach, which has been done before of course.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767763 .

artists in the corpus), and (iv) computing the ltc variant [6] of TF·IDF based on the described vector space model.

2.1 Stand-alone Approaches

PB: A popularity-based recommender that returns the N items listened to most frequently by the entirety of users in the dataset, irrespective of time.

CF: A user-based collaborative filtering approach that recommends N items listened to by u's nearest neighbors in terms of listening histories; neighbors are identified by computing the Inner product between $\vec{p}_u$ and $\vec{p}_v$ $\forall\, v \in V$, where V is the set of all users excluding u.[2]

IB: A content-based (instance-based) approach that identifies for each training item in L_u its nearest neighbors via maximizing cosine similarity between $\vec{w}_i$ and $\vec{w}_j$ $\forall\, j \in J$, where J is the set of all items excluding i.

LB: An extension to **CF** in that we consider as nearest neighbors of u only users that are located in the same country as u. Although this seems to be only a rough description of user location, **LB** turned out to frequently improve performance when integrating such a location component in a hybrid approach.

RB: A baseline that randomly picks users and recommends N artists they listened to.

For **CF**, **IB**, and **LB**, we investigate two aggregation functions to fuse the recommendations contributed by the nearest neighbors,[3] in case they are overlapping: arithmetic mean or maximum between the similarity scores of each item that is recommended by more than 1 neighbor.

2.2 Hybrid Approaches

Leaving aside **RB**, we create hybrid systems that integrate combinations of **PB**, **CF**, **IB**, and **LB**. Fusing the results of all recommenders involved in such a hybrid system is performed by rank aggregation applying Borda counts [5] on the similarity scores of items recommended by each involved approach.[4] This fusion technique has already proven successful for multimodal music recommendation [7]. In total, we investigate the following 14 approaches, again, focusing on algorithmic combinations that performed superior in preliminary experiments:
RB, **PB**, $\mathbf{CF_{mean}}$, $\mathbf{CF_{max}}$, $\mathbf{IB_{mean}}$, $\mathbf{IB_{max}}$, $\mathbf{LB_{mean}}$, **PB**+$\mathbf{CF_{mean}}$, **PB**+$\mathbf{CF_{max}}$, **PB**+$\mathbf{IB_{mean}}$, **PB**+$\mathbf{LB_{mean}}$, $\mathbf{CF_{mean}}$+$\mathbf{IB_{mean}}$, $\mathbf{CF_{mean}}$+$\mathbf{LB_{mean}}$, and **PB**+$\mathbf{CF_{mean}}$+$\mathbf{IB_{mean}}$.

3. USER CHARACTERISTICS

In the following, we define the proposed listening-centric user features, we base our experiments on. We also indicate how we split the users in the dataset according to their feature values.

3.1 Diversity

We define diversity of a user's music taste in two ways: based on how often the user listens to each track in her collection on average and based on the distinct number of genre tags in a user's listening profile. The former variant is computed as $D\_PC_u = P_u/|L_u|$ where P_u denotes the total number of playcounts of u and L_u refers to the set of unique items u listened to. We create two user sets `US_D_PC_[l|h]` by splitting users at their median D_PC value of 2.93 into showing low or high taste diversity, respectively. To compute diversity based on the second definition, we first index all *Last.fm* tags assigned to the items in L_u through a dictionary of almost 2,000 genres fetched from *Freebase*. This naturally results in a set of unique genre tags G_u that describes u's music taste. Diversity is then simply computed as $D\_genres_u = |G_u|$. Users are split into 3 categories `US_D_genres_[l|m|h]`, the borders between them being defined at thirds of accumulated D_genres values.

3.2 Mainstreaminess

We also describe users in terms of the degree to which they prefer music items that are currently popular or rather ignore such trends. We hence coin the term "mainstreaminess" and define the respective user feature as follows. We first define the "mainstream" as a distribution of relative item frequencies among the global set of listening events. To this end, a global playcount vector $\vec{p}$ is defined analogously to $\vec{p}_u$ in Section 2, but on the overall set of listening events and normalized by the sum P of all listening events. The mainstreaminess M_u of a user u then relates the user playcounts to these global playcounts: $M_u = \sum_{p_i \in \vec{p}} \sqrt{\frac{P_{up_i}}{P_u} \cdot \frac{P_{p_i}}{P}}$, where P_{up_i} is the frequency user u listens to each item p_i in vector $\vec{p}$, P_u and P_{p_i} is the total playcount of user u and item p_i, respectively. To account for temporal dynamics in the mainstream and measure to which extent users follow current trends, we introduce a time window which we set to 6 months in our experiments.[5] Considering only listening events within the current window t naturally extends the definition of M_u to M_{ut} of user u in time window t. Shifting t along the temporally ordered listening events of user u and computing the arithmetic mean over all M_{ut} values yields again a scalar value describing u's mainstreaminess. For our experiments, users are categorized into 3 classes for each definition (global and 6-month-average): `US_M_global_[l|m|h]` and `US_M_avg_6m_[l|m|h]`; the borders between them are again defined at thirds of accumulated mainstreaminess values.

3.3 Novelty

This feature models the inclination of user u to listen to unknown music. Splitting u's listening history into time slots of again 6 months, we calculate the percentage of new items listened to, i.e. items appearing for the first time in u's listening history. The novelty N_{ut} of u's listening events in time window t is defined as $N_{ut} = \frac{|\{l \in L_{ut} \,\wedge\, l \notin L_{ux} \forall x < t\}|}{|L_{ut}|}$, where L_{ut} is the entirety of items u listened to in time window t, including duplicates, and $l \notin L_{ux} \forall x < t$ denotes all listening events not listened to by u at any time before t. Averaging over all time slots user u was active in, we obtain u's overall novelty score N_u. We again investigate user sets

[2] As a matter of fact, we omit from $\vec{p}(u)$ the items that occur in the test set.

[3] In case of **CF** and **LB**, nearest neighbors refer to users; in case of **IB**, they refer to items.

[4] We also investigated maximum, mean, sum, and product, but Borda outperformed these in the majority of cases in preliminary experiments.

[5] Preliminary analyses of the data evidenced that a time window of 6 months represents a good trade-off between computational complexity and stability, still incorporating important music trends.

categorized at accumulated thirds into low, medium, and high novelty: `US_N_avg_6m_[l|m|h]`.

4. EVALUATION

To investigate if tailoring music recommendations to listeners within a certain group according to our diversity, mainstreaminess, and novelty definitions improves recommendation results, we perform the following experiments.

4.1 Dataset and Experimental Setup

The used dataset covers 191,108,462 listening events by 16,429 active *Last.fm* users, who listened to 9,163,123 unique tracks from 2,372,601 unique albums by 1,140,014 unique artists. The average number of listening events per user is 11,603 ± 7,130. We investigate 14 recommendation methods (7 stand-alone and 7 hybrids, cf. Section 2) and 14 user categories (cf. Section 3). For each combination of recommendation approach and user category, we perform 5-fold cross-validation on a per-user basis. To this end, we split the entire listening history of each user into a training set containing 80% unique artists and a test set containing the remaining 20%, and perform five experiments, iterating over all five permutations.

4.2 Results and Discussion

We compute average precision, recall, and F1-score as performance measures, where we compute the arithmetic mean over all users. We obtain precision at different recall levels by varying the number of recommended artists between 1 and 1000. Note that the highest achievable recall for **PB** and **CF** is roughly 39%, because of the frequent case that some artists are listened to by only a single user, are hence never correctly recommended. In the following tables, we show for all user categories the best performing approaches in terms of precision, recall, and F1-score.[6] Tables 1 and 2 show the results for the two diversity definitions, Tables 3 and 4, respectively, for the global and 6-month-average of the mainstreaminess definition, and Table 5 for the novelty using a 6-month-window. Statistically significant results, identified by employing the Mann-Whitney U test for equal means of samples [9], are marked with asterisks. For comparative reasons, Table 6 shows the performance measures for all approaches and the entire user set. Comparing the results over all user sets and methods, we make the following observations:

- Grouping users according to the introduced features and performing recommendations within these groups generally tends to outperform recommenders working on the entire user set, in particular, in terms of recall (12 out of 14 cases).

- For all user categories, recommendations based on popularity (**PB**) seem to play a crucial role; however, solely when combined with other approaches. The only case where **PB** alone performs roughly on par is for the user group with high global mainstreaminess, which is trivial to explain.

- The content-based recommender (**IB**) alone performs poorly, independent of user set and performance measure. In contrast, an **IB** component is integrated in the top performing recommenders in terms of recall, in 50% of the user groups. When combined with other methods, it seems that **IB** nicely diversifies the recommendations. This becomes especially evident for users with medium and high inclination to explore novel music, in contrast to low-novelty-users for whom **IB** does not improve results.

- While in general no substantial difference in results could be identified between the two mainstreaminess definitions (`M_global` and `M_avg_6m`), all investigated recommendation methods underperform for users with low mainstreaminess, most of which not even beat the **RB** baseline.

- The genre-based diversity definition (`D_genres`) outperforms the playcount-based one (`D_PC`) in terms of F1-score, except for very low levels of diversity. By far the highest recall levels could be achieved for users with low diversity.

- Integrating location information (**LB**) only yields an improvement for users with a high global mainstreaminess. In combination with the popularity-based recommender (**PB**), these users seem to largely enjoy music currently popular among other users in the same country.

[6]Results for all approaches are available on request.

Table 1: Mean average precision, recall, and F-score for (significantly) best performing stand-alone and hybrid methods on user categories `US_D_PC`.

`US_D_PC_h`			
Method	**Precision**	**Recall**	**F-score**
RB	1.34	6.35	1.85
$PB + CF_{max}$	3.80	17.38	$*$**5.82**
$PB + LB_{mean}$	2.86	$*$**18.91**	4.46
$CF_{mean} + IB_{mean}$	$*$**6.18**	2.43	2.80
`US_D_PC_l`			
Method	**Precision**	**Recall**	**F-score**
RB	1.78	5.05	2.36
CF_{max}	**10.23**	2.33	2.95
$PB + CF_{mean} + IB_{mean}$	3.54	$*$**23.30**	**5.60**

Table 2: Results for user categories `US_D_genres`.

`US_D_genres_h`			
Method	**Precision**	**Recall**	**F-score**
RB	2.62	2.95	1.94
$PB + CF_{max}$	$*$**9.83**	4.93	4.99
$PB + CF_{mean} + IB_{mean}$	6.08	$*$**13.65**	$*$**7.75**
`US_D_genres_m`			
Method	**Precision**	**Recall**	**F-score**
RB	1.86	2.29	1.38
CF_{max}	$*$**6.43**	4.10	3.92
$PB + CF_{max}$	4.18	15.84	$*$**6.50**
$PB + CF_{mean} + IB_{mean}$	3.28	$*$**21.06**	5.56
`US_D_genres_l`			
Method	**Precision**	**Recall**	**F-score**
RB	0.76	2.83	0.84
CF_{max}	**4.17**	8.12	3.65
$PB + IB_{mean}$	1.32	$*$**31.20**	2.50
$PB + CF_{mean}$	3.32	16.03	$*$**4.41**

5. CONCLUSIONS AND OUTLOOK

We proposed three computational features that describe a user's music taste: diversity, mainstreaminess, and novelty. We performed first experiments investigating to which

Table 3: Results for user categories US_M_global.

US_M_global_h			
Method	**Precision**	**Recall**	**F-score**
RB	3.78	8.35	4.86
PB	10.35	$*$**14.62**	7.59
CF_{max}	$*$**13.28**	3.90	5.61
$PB + LB_{mean}$	8.72	13.20	$*$**9.66**
US_M_global_m			
Method	**Precision**	**Recall**	**F-score**
RB	1.37	2.58	1.21
$PB + IB_{mean}$	2.00	$*$**26.00**	3.58
$PB + CF_{mean}$	4.93	12.70	$*$**5.12**
$PB + CF_{mean} + IB_{mean}$	$*$**8.25**	0.86	1.48
US_M_global_l			
Method	**Precision**	**Recall**	**F-score**
RB	**5.15**	5.83	**4.56**
$PB + LB_{mean}$	1.74	**10.67**	2.25

Table 4: Results for user categories US_M_avg_6m.

US_M_avg_6m_h			
Method	**Precision**	**Recall**	**F-score**
RB	3.60	3.82	2.54
CF_{max}	$*$**10.14**	10.94	6.27
$PB + CF_{max}$	7.53	$*$**18.73**	$*$**9.93**
US_M_avg_6m_m			
Method	**Precision**	**Recall**	**F-score**
RB	1.50	5.79	2.18
PB	**5.93**	10.07	3.26
$PB + CF_{max}$	3.56	$*$**18.89**	5.51
$PB + CF_{mean} + IB_{mean}$	5.03	8.90	**5.55**
US_M_avg_6m_l			
Method	**Precision**	**Recall**	**F-score**
RB	**3.02**	7.83	**3.42**
$PB + LB_{mean}$	2.23	**14.77**	2.87

Table 5: Results for user categories US_N_avg_6m.

US_N_avg_6m_h			
Method	**Precision**	**Recall**	**F-score**
RB	1.12	4.57	1.30
PB	**3.89**	10.28	**3.40**
$PB + CF_{mean} + IB_{mean}$	3.51	$*$**11.96**	3.30
US_N_avg_6m_m			
Method	**Precision**	**Recall**	**F-score**
RB	1.78	2.66	1.34
CF_{max}	$*$**6.58**	4.96	3.73
$PB + CF_{mean} + IB_{mean}$	3.60	$*$**23.31**	$*$**5.83**
US_N_avg_6m_l			
Method	**Precision**	**Recall**	**F-score**
RB	3.66	6.54	3.51
PB	**4.27**	9.48	3.91
$PB + CF_{mean}$	4.00	12.64	**3.97**
$PB + LB_{mean}$	3.66	**13.38**	3.97

Table 6: Results for all approaches on the entire user base.

Method	**Precision**	**Recall**	**F-score**
RB	1.55	4.49	1.58
PB	4.82	11.83	4.26
CF_{mean}	5.41	9.26	4.42
CF_{max}	5.22	9.03	4.20
IB_{mean}	1.82	5.09	1.45
IB_{max}	0.44	3.25	0.55
LB_{mean}	2.59	4.74	2.22
$PB + IB_{mean}$	3.95	11.30	3.71
$PB + CF_{mean}$	**6.35**	10.00	**5.14**
$PB + CF_{max}$	6.34	9.70	4.96
$PB + LB_{mean}$	4.58	9.38	4.02
$CF_{mean} + IB_{mean}$	4.51	10.74	4.13
$CF_{mean} + LB_{mean}$	4.91	8.52	4.01
$PB + CF_{mean} + IB_{mean}$	4.99	**13.85**	4.75

extent music recommendations can be tailored when categorizing users according to these features. Comparative experiments on a variety of stand-alone and hybrid recommendation algorithms identified several algorithms that should be preferred when addressing certain user groups. Summarizing our findings, we conclude that hybrid methods generally outperform stand-alone recommenders, irrespective of user group. For users with high mainstreaminess, a popularity component is particularly important, ideally coupled with location-based filtering. Content-based recommenders should not be used as stand-alone, but recommendations benefit when they are integrated with collaborative filtering or popularity recommenders.

While most results may not seem too surprising, the presented experiments serve as a good starting point for further investigations on tailoring music recommendations according to user characteristics. We noticed that the used dataset is biased towards frequent users of *Last.fm*. We will thus assess if results generalize to less active listeners. In addition, we plan to directly integrate diversity, mainstreaminess, and novelty as user factors into matrix factorization algorithms.

Acknowledgments

This research is supported by the EU-FP7 project no. 601166 and by the Austrian Science Fund (FWF): P25655.

6. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. *Recommender Systems Handbook*, chapter Context-Aware Recommender Systems, pages 217–253. Springer, 2011.

[2] L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, K.-H. Lüke, and R. Schwaiger. InCarMusic: Context-Aware Music Recommendations in a Car. In *Proc. EC-Web*, 2011.

[3] O. Celma. *Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

[4] Z. Cheng and J. Shen. Just-for-Me: An Adaptive Personalization System for Location-Aware Social Music Recommendation. In *Proc. ICMR*, Glasgow, UK, April 2014.

[5] J.-C. de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

[6] F. Debole and F. Sebastiani. Supervised Term Weighting for Automated Text Categorization. In *Proc. SAC*, 2003.

[7] M. Kaminskas, F. Ricci, and M. Schedl. Location-aware Music Recommendation Using Auto-Tagging and Hybrid Matching. In *Proc. RecSys*, Hong Kong, China, 2013.

[8] H.-S. Park, J.-O. Yoo, and S.-B. Cho. A context-aware music recommendation system using fuzzy bayesian networks with utility theory. In *FSKD. LNCS (LNAI)*, pages 970–979. Springer, 2006.

[9] B. Prajapati, M. Dunne, and R. Armstrong. Sample Size Estimation and Statistical Power Analyses. *Optometry Today*, 16(7), 2010.

[10] M. Schedl, E. Gómez, and J. Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2–3):127–261, 2014.

[11] M. Schedl, D. Hauger, K. Farrahi, and M. Tkalčič. On the Influence of User Characteristics on Music Recommendation. In *Proc. ECIR*, Vienna, Austria, 2015.

[12] Y. Shi, M. Larson, and A. Hanjalic. Collaborative Filtering Beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.*, 47(1):3:1–3:45, May 2014.

[13] X. Wang, D. Rosenblum, and Y. Wang. Context-aware mobile music recommendation for daily activities. In *Proc. ACM Multimedia*, 2012.

[14] Yuan Cao Zhang, Diarmuid O Seaghdha, Daniele Quercia, Tamas Jambor. Auralist: Introducing Serendipity into Music Recommendation. In *Proc. WSDM*, 2012.