On the Influence of User Characteristics on Music Recommendation Algorithms

Markus Schedl[1], David Hauger[1], Katayoun Farrahi[2], Marko Tkalčič[1]

[1]Department of Computational Perception, Johannes Kepler University, Linz, Austria
[2]Department of Computing, Goldsmith's University of London, UK

Abstract. We investigate a range of *music recommendation algorithm* combinations, *score aggregation functions*, *normalization techniques*, and *late fusion techniques* on approximately 200 million listening events collected through *Last.fm*. The overall goal is to identify superior combinations for the task of artist recommendation. Hypothesizing that user characteristics influence performance on these algorithmic combinations, we consider specific user groups determined by age, gender, country, and preferred genre. Overall, we find that the performance of music recommendation algorithms highly depends on user characteristics.

1 Introduction

Music recommendation within the field of recommender systems is becoming increasingly important since the advent of music streaming platforms that provide access to tens of millions of tracks. At the same time, listeners reveal a lot of personal information in social media, which might play an important role on the quality of music recommendations. However, the relationship between user characteristics and quality of music recommendations has not been thoroughly explored. In this paper, we provide an analysis of various combinations of *recommendation algorithms*, *score aggregation functions*, *normalization techniques*, and *late fusion techniques* on a dataset of almost 200 million listening events from *Last.fm*. Hypothesizing that age, gender, country, and preferred genre influence the quality of recommendations, we further group users according to these aspects and evaluate performance on the resulting user groups.

This work is organized as follows. Section 2 overviews related work. In Sections 3 and 4, we present the aspects we categorize users into and the recommendation models and settings we investigate, respectively. We introduce the dataset used for the experiments, explain the experimental setup, and analyze results in Section 5, before concluding in Section 6.

2 Related Work

Current work on music recommender systems typically employs the same recommendation algorithm to serve different user groups. While it can be argued that matrix factorization techniques may take into account various user aspects, such as temporal dynamics, they still use a single algorithm [4]. In contrast, in this work, we assess how different algorithmic variants of music recommenders perform for different groups of users. In the same vein, Farrahi et al. [5] analyze how aspects of listening frequency, diversity, and mainstreaminess influence recommendation models, but they use a relatively small and sparse dataset mined from microblogs.

Work that integrates user aspects into music recommendation algorithms includes Kaminskas et al. [7], who propose a hybrid matching method to recommend music for places of interest. Baltrunas et al. [1] target music recommendation in a car, taking into account driver and traffic conditions. Zangerle et al. [10] propose an approach that exploits user-based co-occurrences of music items mined from *Twitter* data. Chen and Shen [2] propose a recommendation approach that integrates user location, listening history, music descriptors, and global music popularity trends inferred from microblogs.

In this work, we chose the music platform *Last.fm* to gather a real-world dataset, since it has been shown to attract users of a wide variety of music tastes [9]. In contrast, existing work commonly makes use of rather small and noisy datasets, typically gathered from *Twitter* and including a maximum of a few million listening events [6].

3 User Characteristics

To investigate which aspects of the listener influence the performance of music recommendation algorithms, we categorize each user according to the following attributes. Typewriter font is used to indicate the abbreviations for categories used to indicate user sets for the results.

Age: Listeners from 8 possible age groups are considered. These ranges are [6-17], [18-21], [22-25], [26-30], [31-40], [41-50], [51-60], [61-100]: `US_age_[Start-End]`.

Gender: A listener's gender is considered (i.e. male or female): `US_gender_[male|female]`.

Country: *Last.fm* provides the user with a choice of 240 countries to select from. For reasons of computational complexity and significance of results, we focus on users from the top 6 countries (USA, UK, Brazil, Russia, Germany, and Poland). Each of these has more than 500 users and all other countries are assigned less than half of the number of users of any top 6 country: `US_country_[US|UK|BR|RU|DE|PL]`.

Genre: We categorize listeners according to their preferred genre(s). Assuming that people are typically highly affine to at most 3 different genres, we compute the share of a user's listening events for each genre among all her listening events. Each user is then categorized into all genre classes for which her listening share exceeds 30% of her total listening events. This way a user is assigned none, one, two, or three genre classes. We finally create user sets for 5 representative genres: `US_genre_[jazz|rap|folk|blues|classical]`.

4 Recommendation Methods

We assess several recommendation algorithms for the task of music artist recommendation, in particular, standard *user-based collaborative filtering* (CF), a *popularity-based* algorithm (PB), and an algorithm based on user distance with respect to political or *cultural regions* (CULT). The PB algorithm recommends the most popular artists (i.e. most frequently played) in the dataset. The CULT method defines the target user's nearest neighbors as those that reside in the same country, and recommends their preferred music. As baseline, we include a recommender that proposes items of randomly picked users (RB). For the CF and CULT algorithms, we define two *aggregation functions* (arithmetic mean and maximum) which are used to create an overall ranking of artists to recommend, as an aggregation of similarity scores of the target user's nearest neighbors.

In addition to single methods (PB, $\mathsf{CF}_{[\text{mean,max}]}$, $\mathsf{CULT}_{[\text{mean,max}]}$, and RB), we analyze combinations of two and three algorithms. More precisely, we look into all possible variants: PB+CF, PB+CULT, CF+CULT, and PB+CF+CULT. For these combined variants, a variety of *normalization functions* (n) and *fusion functions* (f) are defined. We consider four methods to normalize the scores of different recommendation methods before fusing their results: n_{none} indicates no normalization is performed; n_{gauss} refers to Gaussian normalization; n_{sumto1} and n_{maxto1} linearly stretches the scores so that their sum equals 1 or their maximum value equals 1, respectively. After scores have been normalized, the results of individual recommenders can be fused. Five fusion functions are investigated: f_{max}, f_{mean}, f_{sum}, $f_{multiply}$, and f_{borda}. While the former four fuse the scores of the individual recommenders directly, by computing their maximum, arithmetic mean, sum, or product, the latter performs rank aggregation based on Borda count [3]. To facilitate perception of individual experiments, we define a standardized scheme. We use sans-serif font for denominations of experiments. For instance, $\mathsf{PB}+\mathsf{CF}_{mean}+\mathsf{CULT}_{max}$ (n_{gauss},$f_{multiply}$) refers to an experiment in which three algorithms (PB, CF, and CULT) are combined. While the CF recommender employs the mean as aggregation function, CULT employs the maximum. Before fusing the results of the three recommenders by multiplying the item scores, Gaussian normalization is performed.

5 Evaluation

5.1 Dataset

In order to conduct experiments on a large scale, a dataset of almost 200 million listening events has been fetched through the *Last.fm* API.[1] To this end, we select a random subset of 16,429 active users and obtain their listening histories of up to 20,000 listening events. After data cleansing, this eventually yields 191,108,462 listening events to 1,140,014 unique artists. The average number of listening events per user is 11,603 ± 7,130.

5.2 Experimental Setup

We perform 5-fold cross-validation on a per-user basis, i.e. using 80% of each user's listening history for training and 20% for testing. Given the components of one recommendation experiment, there is a total of 1,640 different algorithmic combinations per user set (recommendation model, number of recommended items, aggregation function, normalization function, and fusion technique). The investigated 4 user categories with a total of 21 attributes thus require 34,440 individual runs.

We measure performance in terms of precision, recall, and F-measure, for various numbers [10–1000] of recommended artists. Please note that there exists a natural upper limit for achievable recall, because several artists in the dataset are listened to by only a single user, can hence never be recommended. This upper limit is 38.63% for the entire dataset, not grouping by any user set.

[1] http://www.last.fm/api

Table 1. Average and maximum precision, recall, and F-measure for best performing methods and algorithmic combinations, on categories `US_age` (upper part) and `US_gender` (lower part).

Method	**Precision**		**Recall**		**F-score**	
	avg	max	avg	max	avg	max
`US_age_06-17`						
$RB(n_{none})$	1.44	2.06	7.43	19.81	1.63	2.05
$PB + CF_{mean}(n_{none}, f_{max})$	**4.26**	**8.20**	16.44	**34.87**	4.37	5.61
$PB + CF_{mean}(n_{none}, f_{borda})$	4.21	7.41	**16.93**	34.76	**4.42**	**5.69**
`US_age_18-21`						
$RB(n_{none})$	1.51	1.94	5.45	14.37	1.64	2.28
$CF_{mean}(n_{none})$	5.15	**9.66**	14.47	33.02	4.94	6.04
$PB + CF_{mean}(n_{none}, f_{borda})$	**5.37**	8.92	**15.36**	**33.41**	**5.27**	**6.46**
`US_age_22-25`						
$RB(n_{none})$	1.61	1.98	4.60	11.98	1.65	2.37
$PB(n_{none})$	***5.25***	9.02	***11.61***	**25.20**	***4.74***	**5.98**
$CF_{mean}(n_{none})$	4.93	**9.85**	8.34	21.90	4.10	5.74
`US_age_26-30`						
$RB(n_{none})$	1.62	1.95	3.85	10.23	1.59	2.36
$PB(n_{none})$	***5.46***	8.77	***10.24***	**22.41**	***4.73***	5.95
$CF_{mean}(n_{none})$	5.09	**8.97**	9.57	22.30	4.32	**5.27**
`US_age_31-40`						
$RB(n_{none})$	1.71	1.85	3.35	8.92	1.59	2.51
$PB(n_{none})$	***5.90***	**9.93**	***9.18***	**20.20**	***4.72***	**5.88**
`US_age_41-50`						
$RB(n_{none})$	1.79	2.30	3.48	9.38	1.62	2.65
$CF_{mean}(n_{none})$	**6.07**	**9.68**	**9.53**	**20.61**	***4.84***	**6.18**
`US_age_51-60`						
$RB(n_{none})$	1.85	2.36	3.69	9.40	1.68	2.56
$CF_{mean}(n_{none})$	**6.02**	**10.78**	**9.64**	**20.12**	**4.74**	**6.14**
`US_age_61-`						
$RB(n_{none})$	1.45	1.67	3.65	8.75	1.42	2.30
$CF_{mean}(n_{none})$	4.23	**7.51**	8.47	18.74	3.55	**4.43**
$PB + CF_{mean}(n_{none}, f_{max})$	**4.24**	7.51	8.52	18.88	**3.56**	4.43
$PB + CF_{mean}(n_{none}, f_{borda})$	3.87	5.63	**8.59**	**19.37**	3.47	4.27

`US_gender_male`						
$RB(n_{none})$	0.74	1.54	1.70	8.22	0.77	2.10
$PB(n_{none})$	**2.47**	**6.64**	**4.92**	**19.88**	**2.45**	**5.51**
$PB + CF_{mean}(n_{sumto1}, f_{max})$	0.79	6.34	0.79	6.34	0.79	**6.34**
`US_gender_female`						
$RB(n_{none})$	1.78	2.13	5.31	13.87	1.85	2.70
$PB(n_{none})$	**5.63**	9.28	***12.88***	**27.72**	***5.18***	6.47
$PB + CF_{mean}(n_{sumto1}, f_{max})$	3.03	**9.86**	1.52	6.62	1.66	**6.62**

5.3 Discussion

Due to space limitations, we cannot provide here the entire set of results for each algorithmic combination. We hence only show the results of the best performing variants (in terms of average and maximum precision, recall, and F-measure) for each user category and attribute. Results for categories age and gender are shown in Table 1; results for country and genre are depicted in Table 2.

Main general findings from these results are that (i) fusing scores of different recommenders frequently outperforms single variants, (ii) using the mean as aggregation function for CF almost always outperforms the maximum,[2] and (iii) recommendations are overall better when categorizing users according to age and country than according to gender or genre. Analyzing the results per category in detail, we make other interesting observations:

- Younger people seem to be easier to satisfy by recommending overall popular music, whereas mid-aged and elder listeners (41-100) should be offered collaborative filtering recommendations (or combinations that include CF).
- By recommending music using the PB approach it seems slightly easier to satisfy women than men; otherwise no substantial differences between genders can be made out.

[2] This is not the case for currently investigated content-based recommenders.

Table 2. Average and maximum precision, recall, and F-measure for best performing methods and algorithmic combinations, on categories `US_country` (upper part) and `US_genre` (lower part).

Method	**Precision**		**Recall**		**F-score**	
	avg	max	avg	max	avg	max
`US_country_US`						
$RB(n_{none})$	2.00	2.58	4.63	11.93	1.94	2.81
$CF_{mean}(n_{none})$	6.12	**10.93**	11.50	26.57	5.21	6.31
$PB + CF_{max}(n_{sumto1}, f_{sum})$	3.12	6.72	10.94	**27.34**	4.11	**6.96**
$PB + CF_{max}(n_{none}, f_{borda})$	**6.34**	10.46	**12.21**	27.24	**5.52**	6.85
`US_country_UK`						
$RB(n_{none})$	2.11	2.47	4.89	12.61	2.07	3.00
$CF_{mean}(n_{none})$	***6.79***	**12.07**	**12.20**	**26.92**	***5.67***	**7.10**
`US_country_BR`						
$RB(n_{none})$	1.93	2.75	7.37	18.18	2.07	2.74
$CF_{mean}(n_{none})$	***6.44***	**12.35**	**19.41**	**42.59**	***6.30***	**7.87**
`US_country_RU`						
$RB(n_{none})$	1.28	1.65	3.44	9.08	1.26	1.83
$PB(n_{none})$	***4.79***	**8.25**	***10.18***	**21.97**	***4.16***	**5.13**
`US_country_DE`						
$RB(n_{none})$	1.58	1.79	4.06	10.63	1.54	2.29
$CF_{mean}(n_{none})$	***5.73***	**10.16**	**11.94**	**26.79**	***4.96***	**6.18**
`US_country_PL`						
$RB(n_{none})$	1.64	1.96	5.81	14.97	1.77	2.46
$CF_{mean}(n_{none})$	***5.68***	**10.70**	**15.16**	**34.17**	***5.34***	**6.63**
`US_genre_jazz`						
$RB(n_{none})$	1.21	1.49	9.56	26.64	1.47	1.82
$PB(n_{none})$	**2.78**	**5.01**	**16.74**	35.44	**3.13**	**3.88**
$PB + CF_{mean}(n_{none}, f_{multiply})$	2.71	4.92	16.10	**35.63**	3.01	3.76
`US_genre_rap`						
$RB(n_{none})$	0.88	1.00	9.03	25.77	1.17	1.57
$CF_{mean}(n_{none})$	2.58	**5.24**	16.20	33.28	2.90	**3.85**
$PB + CF_{mean}(n_{none}, f_{max})$	**2.59**	5.24	16.42	34.73	**2.90**	3.85
$PB + CF_{mean}(n_{none}, f_{multiply})$	2.22	3.73	16.78	**36.67**	2.66	3.48
$PB + CF_{mean}(n_{none}, f_{borda})$	2.48	4.77	**17.21**	35.94	2.87	3.60
`US_genre_folk`						
$RB(n_{none})$	1.15	1.50	9.12	25.53	1.46	1.94
$CF_{mean}(n_{none})$	3.57	**7.42**	18.41	38.10	3.86	5.10
$PB + CF_{mean}(n_{none}, f_{multiply})$	3.18	5.74	18.44	**39.55**	3.59	4.58
$PB + CF_{mean}(n_{none}, f_{borda})$	3.46	7.05	**18.99**	39.19	3.82	4.96
$PB + CF_{mean} + CULT_{mean}(n_{none}, f_{max})$	**3.57**	7.42	18.56	38.86	**3.87**	**5.10**
`US_genre_blues`						
$RB(n_{none})$	1.59	2.88	6.66	23.99	1.77	3.18
$PB + CF_{mean}(n_{maxto1}, f_{mean})$	2.73	4.73	**24.20**	**53.88**	3.30	4.10
$PB + CF_{max}(n_{none}, f_{multiply})$	**2.85**	**5.93**	23.11	52.68	**3.32**	3.96
$PB + CF_{mean} + CULT_{mean}(n_{maxto1}, f_{mean})$	2.72	4.67	24.20	53.88	3.30	**4.11**
`US_genre_classical`						
$RB(n_{none})$	1.28	2.35	3.74	11.65	1.27	2.35
$CF_{mean}(n_{none})$	2.29	**7.08**	6.58	16.49	2.18	4.38
$PB + CF_{mean}(n_{maxto1}, f_{mean})$	2.77	5.54	17.31	37.77	**3.13**	4.42
$PB + CF_{mean}(n_{none}, f_{borda})$	2.79	6.23	16.85	37.10	3.08	**4.56**
$PB + CF_{max}(n_{sumto1}, f_{mean})$	**2.81**	6.23	17.07	37.35	3.11	4.52
$PB + CF_{max}(n_{maxto1}, f_{mean})$	2.76	5.38	**17.32**	**37.85**	3.13	4.50

- While listeners in most investigated countries are served well by CF approaches, Russian listeners seem to prefer highly popular mainstream music (PB). For US citizens, the right mixture of popular music and music listened to by like-minded people yields best results.
- Including cultural aspects (CULT) most strongly contributes to increased performance for lovers of folk and blues. The surprisingly good performance of PB for jazz aficionados indicates that they may prefer overall popular jazz music, whereas combinations of PB and CF provide most accurate recommendations for fans of rap and classical music.

In order to see if there is a significant winning method within the user groups, we perform pairwise significance tests between the best method within each group and the other methods within that group. As the distributions of the performance metrics (precision, recall, and F-score) are not normal, we employ the

Mann-Whitney U test for equal medians of two samples [8]. We mark significant results in italics in the results tables.

6 Conclusion and Outlook

The overall finding of our study is that music recommendation can be improved by tailoring to different listener categories. There is no single method that fits everyone; rather a combination of individual recommendation models and variants should be considered for each user group.

We are currently conducting experiments using a larger variety of user-specific factors, including categories related to listening frequency, temporal aspects of music consumption, and openness to unknown music. In the future, we would also like to investigate the influence of personality traits on music recommendation and music taste in general. Song-level recommendation experiments and the related topic of addressing data sparsity, as well as looking into content-based algorithms, constitute other research directions.

7 Acknowledgments

This research is supported by the EU-FP7 project no. 601166 and by the Austrian Science Fund (FWF): P25655. The authors are furthermore most thankful to Johann Messner for providing his expertise in scientific computing.

References

1. 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*, Toulouse, France, Aug–Sep 2011.
2. Z. Cheng and J. Shen. Just-for-Me: An Adaptive Personalization System for Location-Aware Social Music Recommendation. In *Proc. ICMR*, Glasgow, UK, Apr 2014.
3. J.-C. de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.
4. G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. The Yahoo! Music Dataset and KDD-Cup'11. *JMLR: Proceedings of KDD-Cup 2011 competition*, 18:3–18, Oct 2012.
5. K. Farrahi, M. Schedl, A. Vall, D. Hauger, and M. Tkalčič. Impact of Listening Behavior on Music Recommendation. In *Proc. ISMIR*, Taipei, Taiwan, Oct 2014.
6. 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, Nov 2013.
7. M. Kaminskas, F. Ricci, and M. Schedl. Location-aware Music Recommendation Using Auto-Tagging and Hybrid Matching. In *Proc. RecSys 2013*, Hong Kong, China, Oct 2013.
8. B. Prajapati, M. Dunne, and R. Armstrong. Sample Size Estimation and Statistical Power Analyses. *Optometry Today*, 16(7), 2010.
9. M. Schedl and M. Tkalčič. Genre-based Analysis of Social Media Data on Music Listening Behavior. In *Proc. ACM Multimedia Workshop: ISMM*, Orlando, FL, USA, Nov 2014.
10. E. Zangerle, W. Gassler, and G. Specht. Exploiting Twitter's Collective Knowledge for Music Recommendations. In *Proc. WWW Workshop: #MSM*, Apr 2012.