Leveraging the structure of musical preference in content-aware music recommendation

Paul Magron, Cédric Févotte

Leveraging the structure of musical preference in content-aware music recommendation

Paul Magron*, Cédric Févotte*

Abstract

State-of-the-art music recommendation systems are based on collaborative filtering, which predicts a user's interest from his listening habits and similarities with other users' profiles. These approaches are agnostic to the song content, and therefore face the cold-start problem: they cannot recommend novel songs without listening history. To tackle this issue, content-aware recommendation incorporates information about the songs that can be used for recommending new items. Most methods falling in this category exploit either user-annotated tags, acoustic features or deeply-learned features. Consequently, these content features do not have a clear musical meaning, thus they are not necessarily relevant from a musical preference perspective. In this work, we propose instead to leverage a model of musical preference which originates from the field of music psychology. From low-level acoustic features we extract three factors (arousal, valence and depth), which have been shown appropriate for describing musical taste. Then we integrate those into a collaborative filtering framework for content-aware music recommendation. Experiments conducted on large-scale data show that this approach is able to address the cold-start problem, while using a compact and meaningful set of musical features.

Keywords— Content-aware music recommendation, musical preference structure, collaborative filtering, cold-start problem, matrix factorization.

1 Introduction

Music recommendation [1] consists in predicting users' listening habits in order to suggest them novel tracks that they might enjoy. This task, which is at the core of many commercial platforms, has been extensively investigated, but remains challenging due to the complexity of music and to the lack of explicit and reliable users feedback [2]. Among these many challenges, accounting for contextual information is of paramount importance, so that recommendations are adequate with a specific situation or location [3]. Besides, music recommender systems face the *cold-start* problem [4]: when a new song is added to a music catalog, there is no interaction data between the users and this song. Consequently, the system cannot properly recommend this new item to existing users. This problem, which is still considered as a major challenge in music recommendation research [5], is the topic of interest of the present paper.

State-of-the-art approaches for music recommendation are based on collaborative filtering [2], a family of techniques which rely solely on users' listening history: the interest of a given user for a given song is predicted using similarities between various user profiles. The users' feedback are most often implicit and in the form of *playcounts*, that is, how many times a given user has listened to a particular song. This listening history is however noisy and lacks negative feedback data [2]. To alleviate this problem, weighted matrix factorization (WMF) techniques [6, 7] process binarized data computed from the raw playcounts, and associate a measure of *confidence* to these binarized feedback. This family of techniques has shown good performance in music recommender systems [8]. More recently, with the advent of deep learning, collaborative filtering techniques using deep neural networks (DNNs) have been proposed and yield promising results [9, 10]. However, WMF techniques remain a solid choice, as they are light, flexible, and still compete with neural collaborative filtering approaches [11].

Such approaches are however agnostic to any form of song-related content, which becomes a major issue for new items: a collaborative filtering method is not able to recommend a song without listening history, and therefore face the cold-start problem. This problem has been tackled with content-based methods, which aim to exploit additional information about the items for recommendation [12, 13, 14]. For instance, [15] exploits tags in a co-factorization framework, deep content-based approaches [16, 17] use acoustic features in conjunction with collaborative filtering, and [8] uses the last hidden layer of an auto-tagging DNN as content feature.

Even though these approaches have been shown useful for addressing the cold-start problem, these content features lack a clear musical meaning. Consequently, they are not necessarily relevant in terms of musical preference, and appropriate for recommendation. On the other hand, research has been conducted on the structure

*IRIT, Université de Toulouse, CNRS, Toulouse, France (e-mail: firstname.lastname@irit.fr).

of musical preference [18]. Recent studies [19, 20, 21] notably show that musical preference can be described by using a set of three factors termed *arousal*, *valence* and *depth* (AVD). The usefulness of musical preference models for recommendation has been pointed out in [22], but to the best of our knowledge, it has only been exploited in [23]. However, this approach is not based on collaborative filtering and relies on expert ratings for estimating a genre-specific musical preference model, which hinders its deployment at a larger scale.

In this paper, we propose to leverage the AVD model of musical preference in music recommendation based on collaborative filtering. Unlike prior work using tag-based or deeply-learned features, we claim that using descriptors that characterize musical preference is appropriate for recommendation. We therefore aim at bridging the gap between music recommendation and music psychology, which has notably been motivated in a recent survey [5]. We compute the AVD factors and we integrate them as content features into a WMF-based framework for collaborative filtering. We experimentally assess the potential of this method for cold-start recommendation, where it notably outperforms a pure content-based method.

The rest of this paper is structured as follows. Section 2 presents the collaborative filtering framework. Section 3 describes the AVD model. Section 4 details the music recommendation experiments. Finally, Section 5 draws some concluding remarks.

2 Collaborative filtering model

In this section, we present the content-aware collaborative filtering framework [8] on which our method is based. Even though the literature on matrix factorization models is quite abundant, we adopted this WMF framework as it was shown appropriate for handling implicit feedback data [2, 16]. We first describe the baseline WMF model, and then present its content-aware extension.

2.1 Weighted matrix factorization

We consider playcount data $\mathbf{Y} \in \mathbb{R}^{U \times I}$, where U and I respectively denote the number of users and items. The (u,i)-th entry of $\mathbf{Y}$, denoted $y_{u,i}$, is the number of times the song i has been listened to by user u.

WMF [2, 7] models the data as the product of two low-dimensional factors: a (transposed) user preferences matrix $\mathbf{W} \in \mathbb{R}^{K \times U}$ and an item attributes matrix $\mathbf{H} \in \mathbb{R}^{K \times I}$. The rank K of the decomposition is chosen such that $K(U+I) << UI$ to ensure dimensionality reduction. In a probabilistic framework [6], these factors are usually assumed to be drawn from centered normal distributions with scaling parameters λ_W and λ_H:

$$\begin{aligned} \forall u \in \{1,...,U\},\ \mathbf{w}_u &\sim \mathcal{N}(0, \lambda_W^{-1}\mathbf{I}_K), \\ \forall i \in \{1,...,I\},\ \mathbf{h}_i &\sim \mathcal{N}(0, \lambda_H^{-1}\mathbf{I}_K), \end{aligned} \tag{1}$$

where $\mathbf{w}_u$ (resp. $\mathbf{h}_i$) is the u-th (resp. i-th) column of $\mathbf{W}$ (resp. $\mathbf{H}$), and $\mathbf{I}_K$ is the identity matrix of dimension K. In order to better account for the over-dispersed nature of the raw data, it is common to consider the binarized playcount matrix $\mathbf{R}$, which indicates whether a user has listened to a song more than a certain amount of times or not [8]. Note that alternative strategies have been proposed to address this issue, notably by crafting refined statistical models based on the Poisson or compound Poisson distributions [24, 25, 26].
The generative process for the observed data is then:

$$\forall u,\ i,\ r_{u,i} \sim \mathcal{N}(\mathbf{w}_u^\mathsf{T}\mathbf{h}_i, c_{u,i}^{-1}), \tag{2}$$

where $^\mathsf{T}$ denotes the matrix transpose and

$$c_{u,i} = 1 + \alpha \log\left(1 + \frac{y_{u,i}}{\epsilon}\right) \tag{3}$$

is called the *confidence*, with $\alpha = 2$ and $\epsilon = 10^{-6}$ [2, 8]. The model defined by (1) and (2) is estimated in a maximum a posteriori sense. Then, recommendation can be done based on the predicted ratings $\hat{r}_{u,i} = \mathbf{w}_u^\mathsf{T}\mathbf{h}_i$. However, this only applies to songs for which some listening history is available, hence facing the cold-start problem.

2.2 Content-aware WMF

In order to incorporate content information in WMF, the authors in [8] propose to modify the prior on the item attributes matrix (1), which rewrites:

$$\mathbf{h}_i \sim \mathcal{N}(\phi(\mathbf{z}_i), \lambda_H^{-1}\mathbf{I}_K), \tag{4}$$

where $\mathbf{z}_i \in \mathbb{R}^L$ is a content feature vector of dimension L, and ϕ is a mapping between this vector and the item attributes vector $\mathbf{h}_i$. The authors consider a linear mapping: $\phi(\mathbf{z}_i) = \mathbf{B}\mathbf{z}_i$ with $\mathbf{B} \in \mathbb{R}^{K \times L}$. Estimating the

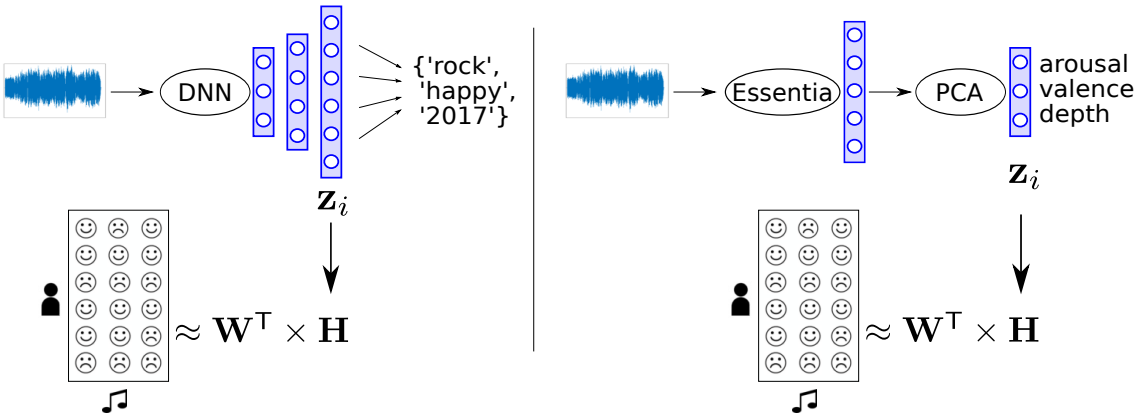

Figure 1: Illustration of the content-aware collaborative filtering: content features can be extracted using an auto-tagging DNN (left, as in [8]) or using the AVD model (right, proposed).

parameters with maximum a posteriori results in the following optimization problem:

$$\min_{\mathbf{W},\mathbf{H},\mathbf{B}} \sum_{u,i} c_{u,i}(r_{u,i} - \mathbf{w}_u^\top \mathbf{h}_i)^2 + \lambda_W \sum_u ||\mathbf{w}_u||^2 + \lambda_H \sum_i ||\mathbf{h}_i - \mathbf{B}\mathbf{z}_i||^2, \quad (5)$$

where $||.||$ denotes the Euclidean norm. Canceling the gradient of the loss in (5) with respect to each parameter iteratively yields the following update rules:

$$\mathbf{w}_u = (\mathbf{H}\text{diag}(\mathbf{c}_u)\mathbf{H}^\top + \lambda_W \mathbf{I}_K)^{-1}\mathbf{H}\text{diag}(\mathbf{c}_u)\mathbf{r}_u, \quad (6)$$

$$\mathbf{h}_i = (\mathbf{W}\text{diag}(\mathbf{c}_i)\mathbf{W}^\top + \lambda_H \mathbf{I}_K)^{-1}(\mathbf{W}\text{diag}(\mathbf{c}_i)\mathbf{r}_i + \lambda_H \mathbf{B}\mathbf{z}_i), \quad (7)$$

$$\mathbf{B} = \mathbf{H}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top + \lambda_B \mathbf{I}_L)^{-1}, \quad (8)$$

where $\mathbf{r}_u = [r_{u,1}, ..., r_{u,I}]^\top$ and $\mathbf{r}_i = [r_{1,i}, ..., r_{U,i}]^\top$ (and similarly for $\mathbf{c}_u$ and $\mathbf{c}_i$), $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_I]$, and λ_B is a small offset added to the diagonal of $\mathbf{Z}\mathbf{Z}^\top$ in order to avoid numerical problems when inverting it [8].

The advantage of this approach is two-fold. First, the content features serve as a regularizer for the collaborative filtering model parameters, and promotes a more meaningful representation learning. Second, this model can be used for solving the cold-start problem. Indeed, for a new item for which there is no user-item interaction, one can still predict a rating using $\hat{r}_{u,i} = \mathbf{w}_u^\top \mathbf{B}\mathbf{z}_i$, from which recommendation can be performed.

3 Content features

In this section, we briefly describe the deep features used in [8]. We then motivate the use of the AVD model which we present subsequently. These two approaches are illustrated in Figure 1.

3.1 Deep features

In [8], the authors use as content features a latent representation that is extracted from an auto-tagging DNN, which is a system that aims at predicting music-related labels (such as genre) from music audio [27]. A set of 1028 vector-quantized timbre features is calculated for each song, and fed as input to a DNN which predicts a set of tags associated with each song among 581 possible tags. The proposed DNN consists of 3 feed-forward layers which serve as feature extractor, plus a logistic regression layer which serves as classifier. Once the tagging network is trained in a supervised fashion, the authors use the output of the last hidden layer as content features $\mathbf{z}_i$ of dimension $L = 1200$.

This approach allows to alleviate the cold-start problem, but it suffers from several drawbacks. First, the quality of the content features is highly dependent on the performance of the auto-tagging network. While these methods have significantly improved in the recent years [27], they are still limited by the noisy nature of the tags, which are user-annotated. Besides, this approach remains computationally demanding and the high dimensionality of the features suggests that it might be strongly redundant. Finally, the extracted content features are not guaranteed to characterize musical preference, and therefore lack a clear musical meaning, which we propose to overcome in the following.

3.2 The AVD model

Studies in the field of psychology of music have been conducted in order to design a model of musical preference. Early research has pointed out that the preference for musical genres can be described in terms of five factors, termed *mellow*, *unpretentious*, *sophisticated*, *intense*, and *contemporary* (MUSIC) [18]. The MUSIC model exhibited high correlation with genre, but later studies [19, 20, 21] focused on more general musical preference attributes (i.e., non-genre specific). These studies outlined that musical preference can be conceptualized using three factors:

- *Arousal*, which describes how energetic and intense a music piece is (e.g., rousing or calm);
- *Valence*, which is related to the perceived emotional content of a song (e.g., happy or sad);
- *Depth*, which is a general indicator of level of sophistication (e.g., complicated or simple).

This AVD model has been shown to correlate well with high-level descriptors of music, whether those are expert-annotated [20] or automatically computed [21]. Since the AVD factors are a compact representation that is meaningful from a musical preference perspective, we claim that they are appropriate content features to incorporate in the collaborative filtering model presented in Section 2.2.

To compute the AVD factors, we follow the procedure used in [21]. For each song, musical features are computed using the Essentia toolbox [28], a collection of a variety of music information retrieval algorithms. In a nutshell, Essentia takes as input the raw audio waveform and outputs a number of features, both low-level (e.g., spectral and temporal descriptors) and high-level (e.g., descriptors such as "happy", "sad" or "danceable"). We retain the 16 high-level features listed in Table 1, which we scale to have zero mean and unit variance. We then perform a principal component analysis (PCA) with 3 factors and oblimin rotation, which allows the factors to be non-orthogonal for better interpretability [29]. The resulting principal components are the AVD factors, which we use as content features $\mathbf{z}_i$ of dimension $L = 3$.

Note that we also considered a larger set of Essentia features comprising both low- and high-level descriptors (84 in total). This yields principal components with overall less correlation with high-level features. This suggests that the AVD model mostly captures the information contained in high-level features, and that adding some extra low-level information hampers the easiness to interpret these factors. Finally, this approach did not yield any significant improvement in terms of recommendation, thus we do not report hereafter the detailed results obtained in this setting.

4 Experiments

In this section, we present the experiments conducted to assess the potential of the AVD model for music recommendation. In the spirit of reproducible research, we provide the code related to these experiments,[1] as well as the Essentia features data.[2]

4.1 Protocol

Dataset. We use the Taste Profile dataset which is part of the Million Song Dataset [30]. It provides listening counts of 1 million users and 380,000 songs. We only keep the songs whose Essentia (and consequently AVD) features can be calculated, which corresponds to a total of 204,316 songs [21]. The playcount data is binarized by retaining values of five or higher as implicit feedback [31]. As in [8, 15], in order to keep the computational burden low, we retain the top songs and users (sorted by playcounts) and we remove inactive users and items (that is, we only keep users who listened to at least 20 songs, and songs which have been listened to by at least 50 users). The resulting dataset comprises 9,132 users and 7,674 songs, for a total of 247,414 (0.35 %) non-zero playcounts.

We use 95 % of the songs for training the WMF model, which corresponds to an *in-matrix* recommendation task. That is, the corresponding playcounts are split into 70 %, 20 % and 10 % for training, validation and in-matrix testing, respectively. The playcounts corresponding to the remaining 5 % of the songs are used for an *out-of-matrix* recommendation task. This corresponds to the cold-start scenario, since the WMF model has not been trained on these songs.

Methods. We compare the content-free WMF (used as a baseline for in-matrix recommendation) to its content-aware counterpart. As for content features, we use the AVD factors estimated according to the protocol described in Section 3.2, as well as the Essentia features before PCA. Even though the deep features presented in Section 3.1 would constitute an interesting comparison reference, we were not able to reproduce the results from [8] as we faced some technical issues when re-implementing the auto-tagging network. Following previous studies [8], WMF is run with 20 iterations and rank $K = 50$. The hyperparameters λ_W and λ_H are tuned on the validation set using the normalized discounted cumulative gain (NDCG, see below), and $\lambda_B = 10^{-2}$.

[1] `https://github.com/magronp/mus-reco-avd`

[2] `https://zenodo.org/record/3860557#.Xs-b7R9fg5l`

Since content-free WMF cannot perform an out-of-matrix recommendation task, we implemented a pure content-based method as a baseline for cold-start recommendation based on [23]. This method consists in computing a mean AVD preference vector for each user from the training set, and then performing recommendations based on similarities between this mean AVD preference vector and the AVD factors extracted from novel songs. Note that the original method in [23] used the MUSIC factors [18], but we use here the AVD factors for a fair comparison with our proposed content-aware collaborative filtering technique.

Evaluation metric. We use the NDCG [32] as a measure of the overall quality of the recommendation. For each user u, we compute a ranked list of items in the test set based on the predicted preference $\hat{\mathbf{r}}_u$. We then compute the relevance of this list with respect to the ground truth preferences (that is, the ratings that were left out for testing): $\text{rel}_{u,i} = 1$ if the item i is in the listening history of user u and 0 otherwise. In order to favor recommendations that place the test items high in the list, we apply a discounted weight to the relevance, which yields the discounted cumulative gain (DCG), and from which its normalized version (ranging from 0 to 1) can be obtain:

$$\text{DCG}_u = \sum_{i=1}^{I} \frac{\text{rel}_{u,i}}{\log_2(i+1)}, \; \text{NDCG}_u = \frac{\text{DCG}_u}{\text{IDCG}_u}, \tag{9}$$

where IDCG is the ideal DCG, which corresponds to the DCG of a perfectly ranked list. Finally, these scores are averaged over users to yield an overall recommendation performance.

4.2 Features analysis

First, we analyze the AVD factors computed according to the protocol presented in Section 3.2. We report in Table 1 the correlation between the high-level Essentia features and the AVD factors. We observe that these correlations are consistent with the definition of these factors. Indeed, the arousal component positively correlates with features such as "rousing" or "aggressive" and negatively correlates with "relaxed" or "sad". The valence factor has a high loading on features such as "happy", "fun" and "danceable", while it is negatively correlated with "aggressive" and "intense". In order to subjectively assess the validity of these factors, we examine the songs with maximum and minimum AVD values (the interested reader can use the provided code to replicate and extend this analysis). We observe that maximum arousal songs belong to the punk, rock or metal genres, and exhibit relatively high aggressiveness and energy, while low-valence pieces are mostly tracks with a dark and/or sad atmosphere and lyrical content. These subjective assessments are consistent with the correlations between high-level Essentia features and the AVD factors highlighted in Table 1.

The meaning of the depth factor is less clear: it highly correlates with "electronic" and its lowest loading is on the "tonal" feature, which suggests that this factor may characterize contemporary and intricate pieces, which are indeed encountered among the top-depth songs. However, its high correlation with the "danceable" feature might appear as contradictory with this description. Note that this ambiguity of the depth factor, as well as these general trends, are overall consistent with the findings of previous studies [20, 21].

4.3 Recommendation results

Table 2 presents the results in terms of NDCG for in- and out-of-matrix recommendation. For in-matrix recommendation, the performance of the methods are similar. A slight improvement is obtained when incorporating some content information within WMF in the form of Essentia features, but this improvement is no longer observed when using the AVD model instead. This framework is known to work quite well on those datasets [8], thus this collaborative filtering method do not benefit greatly from adding extra content information for recommending items which already have some listening history.

The advantage of content-aware WMF appears more clearly for out-of-matrix recommendation, where the usage of the proposed features alleviates the cold-start problem. Using the set of Essentia features slightly outperforms using the AVD factors. Overall, both content-aware WMF methods outpeform the pure content baseline. This reveals the appropriateness of the AVD model for addressing the cold-start problem, and it particularly outlines its relevance when combined with user-iterm interaction data in a collaborative filtering-based framework.

Finally, unlike the approach in [8] which used deep features, our approach is light (it does not require to train an auto-tagging DNN) and relies on content features with a clear musical meaning.

5 Conclusion

In this paper, we proposed to leverage the AVD model of musical preference for music recommendation, which was shown to correlate well with high-level musical properties. We extracted the AVD factors from audio excerpts and integrated them into a content-aware collaborative filtering method. This approach has shown its potential for music recommendation, notably for addressing the cold-start problem in a computationally very light framework.

Table 1: Correlations between the high-level Essentia features and the AVD factors.

	Arousal	Valence	Depth
Mirex clusters			
1 (Rousing, Passionate)	**0.56**	-0.11	-0.12
2 (Fun, Cheerful)	-0.03	**0.78**	0.01
4 (Humorous, Witty)	-0.05	**0.63**	0.12
5 (Aggressive, Intense)	0.34	**-0.55**	0.30
Mood-related			
Aggressive	**0.63**	-0.48	-0.02
Happy	**0.52**	0.37	-0.36
Party	**0.69**	0.05	0.40
Relaxed	**-0.84**	0.01	0.14
Sad	**-0.80**	0.07	-0.21
Sound-related			
Acoustic	**-0.78**	0.04	-0.25
Average loudness	**0.59**	0.14	-0.07
Danceable	0.23	0.42	**0.52**
Dissonance	**0.86**	-0.03	-0.04
Dynamic complexity	**-0.57**	0.07	0.21
Electronic	0.08	-0.01	**0.74**
Instrumental	-0.35	-0.06	0.23
Tonal	0.04	0.15	**-0.60**

Table 2: Recommendation performance expressed with the NDCG averaged over users (higher is better).

	In-matrix	Out-of-matrix
Content-free WMF	0.35	–
Pure content	–	0.19
Content-aware WMF		
Essentia	0.36	0.22
AVD	0.35	0.21

Future work will focus on combining this model with other types of content (such as tags or lyrics) in order to fully exploit the available data for recommendation. Finally, we will also explore alternative mappings between content features and item attributes, such as non-linear and/or deep.

6 Acknowledgments

This work is supported by the European Research Council (ERC FACTORY-CoG-6681839). We thank K. R. Fricke, D. M. Greenberg and P. J. Rentflow for fruitful discussion about the AVD model and for providing the data used in this work. We also thank Olivier Gouvert for his insightful comments and for proof-reading the paper.

References

[1] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, "Music recommender systems," in *Recommender Systems Handbook*, pp. 453–492. Springer, 2015.

[2] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. IEEE International Conference on Data Mining (ICDM)*, December 2008.

[3] M. Gillhofer and M. Schedl, "Iron Maiden while jogging, Debussy for dinner? an analysis of music listening behavior in context," in *Proc. International conference on MultiMedia Modeling (MMM)*, January 2015.

[4] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proc. International Conference on Research and Development in Information Retrieval (SIGIR)*, August 2002.

[5] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 95–116, June 2018.

[6] A. Mnih and R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. International Conference on Neural Information Processing Systems (NIPS)*, December 2007.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, August 2009.

[8] D. Liang, M. Zhan, and D. Ellis, "Content-aware collaborative music recommendation using pre-trained neural networks," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, October 2015.

[9] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. International World Wide Web Conference (WWW)*, April 2017.

[10] W. Chen, F. Cai, H. Chen, and M. De Rijke, "Joint neural collaborative filtering for recommender systems," *ACM Transactions on Information Systems*, vol. 37, no. 4, August 2019.

[11] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? a worrying analysis of recent neural recommendation approaches," in *Proc. ACM Conference on Recommender Systems (RecSys)*, 2019.

[12] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, October 2006.

[13] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2011.

[14] P. K. Gopalan, L. Charlin, and D. Blei, "Content-based recommendations with Poisson factorization," in *Advances in Neural Information Processing Systems*, December 2014.

[15] O. Gouvert, T. Oberlin, and C. Févotte, "Matrix co-factorization for cold-start recommendation," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, September 2018.

[16] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in Neural Information Processing Systems*, December 2013.

[17] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proc. ACM International Conference on Multimedia*, November 2014.

[18] P. J. Rentfrow, L. R. Goldberg, and D. J. Levitin, "The structure of musical preferences: A five-factor model," *Journal of Personality and Social Psychology*, vol. 100, no. 6, pp. 1139–1157, June 2011.

[19] D. M. Greenberg, M. Kosinski, D. J. Stillwell, B. L. Monteiro, D. J. Levitin, and P. J. Rentfrow, "The song is you: Preferences for musical attribute dimensions reflect personality," *Social Psychological and Personality Science*, vol. 7, no. 6, pp. 597–605, August 2016.

[20] K. R. Fricke, D. M. Greenberg, P. J. Rentfrow, and P. Y. Herzberg, "Computer-based music feature analysis mirrors human perception and can be used to measure individual music preference," *Journal of Research in Personality*, vol. 75, pp. 94 – 102, August 2018.

[21] K. R. Fricke, D. M. Greenberg, P. J. Rentfrow, and P. Y. Herzberg, "Measuring musical preferences from listening behavior: Data from one million people and 200,000 songs," *Psychology of Music*, September 2019.

[22] A. Laplante, "Improving music recommender systems: What can we learn from research on music tastes?," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, October 2014.

[23] M. Soleymani, A. Aljanaki, F. Wiering, and R. C. Veltkamp, "Content-based music recommendation using underlying music preference structure," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, June 2015.

[24] P. K. Gopalan, J. M. Hofman, and D. Blei, "Scalable recommendation with hierarchical Poisson factorization," in *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2014.

[25] M. Basbug and B. Engelhardt, "Hierarchical compound Poisson factorization," in *Proc. International Conference on Machine Learning (ICML)*, June 2016.

[26] O. Gouvert, T. Oberlin, and C. Févotte, "Recommendation from raw data with adaptive compound Poisson factorization," in *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2019.

[27] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, September 2018.

[28] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: An audio analysis library for music information retrieval," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, November 2013.

[29] I. T. Jolliffe, *Principal Component Analysis*, p. 488, Springer-Verlag New York, 2020.

[30] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, October 2011.

[31] V.-A. Tran, R. Hennequin, J. Royo-Letelier, and M. Moussallam, "Improving collaborative metric learning with efficient negative sampling," in *Proc. International Conference on Research and Development in Information Retrieval (SIGIR)*, July 2019.

[32] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, and W. Chen, "A theoretical analysis of NDCG type ranking measures," in *Proc. Conference on Learning Theory (COLT)*, April 2013.