Current Challenges and Visions in Music Recommender Systems Research

Markus Schedl
Johannes Kepler University
Linz, Austria
markus.schedl@jku.at

Hamed Zamani
University of Massachusetts
Amherst, USA
zamani@cs.umass.edu

Ching-Wei Chen
Spotify USA Inc.
New York City, USA
cw@spotify.com

Yashar Deldjoo
Politecnico di Milano
Milan, Italy
yashar.deldjoo@polimi.it

Mehdi Elahi
Free University of Bozen-Bolzano
Bolzano, Italy
meelahi@unibz.it

ABSTRACT

Music recommender systems (MRS) have experienced a boom in recent years, thanks to the emergence and success of online streaming services, which nowadays make available almost all music in the world at the user's fingertip. While today's MRS considerably help users to find interesting music in these huge catalogs, MRS research is still facing substantial challenges. In particular when it comes to build, incorporate, and evaluate recommendation strategies that integrate information beyond simple user–item interactions or content-based descriptors, but dig deep into the very essence of listener needs, preferences, and intentions, MRS research becomes a big endeavor and related publications quite sparse.

The purpose of this trends and survey article is twofold. We first identify and shed light on what we believe are the most pressing challenges MRS research is facing, from both academic and industry perspectives. We review the state of the art towards solving these challenges and discuss its limitations. Second, we detail possible future directions and visions we contemplate for the further evolution of the field. The article should therefore serve two purposes: giving the interested reader an overview of current challenges in MRS research and providing guidance for young researchers by identifying interesting, yet under-researched, directions in the field.

KEYWORDS

music recommender systems; challenges; automatic playlist continuation; user-centric computing

1 INTRODUCTION

Research in music recommender systems (MRS) has recently experienced a substantial gain in interest both in academia and industry [121]. Thanks to music streaming services like Spotify, Pandora, or Apple Music, music aficionados are nowadays given access to tens of millions music pieces. By filtering this abundance of music items, thereby limiting choice overload [14], MRS are often very successful to suggest songs that fit their users' preferences. However, such systems are still far from being perfect and frequently produce unsatisfactory recommendations. This is partly because of the fact that users' tastes and musical needs are highly dependent on a multitude of factors, which are not considered in sufficient depth in current MRS approaches that are typically centered on the core concept of user–item interactions, or sometimes content-based item descriptors. In contrast, we argue that satisfying the users' musical entertainment needs requires taking into account intrinsic, extrinsic, and contextual aspects of the listeners [2], as well as more decent interaction information. For instance, personality and emotional state of the listeners (intrinsic) [50, 108] as well as their activity (extrinsic) [54, 142] are known to influence musical tastes and needs. So are users' contextual factors including weather conditions, social surrounding, or places of interest [2, 74]. Also the composition and annotation of a music playlist or a listening session reveals information about which songs go well together or are suited for a certain occasion [95, 150]. Therefore, researchers and designers of MRS should reconsider their users in a holistic way in order to build systems tailored to the specificities of each user.

Against this background, in this trends and survey article, we elaborate on what we believe to be amongst the most pressing current challenges in MRS research, by discussing the respective state of the art and its restrictions (Section 2). Not being able to touch all challenges exhaustively, we focus on *cold start*, *automatic playlist continuation*, and *evaluation* of MRS. While these problems are to some extent prevalent in other recommendation domains too, certain characteristics of music pose particular challenges in these contexts. Among them are the short duration of items (compared to movies), the high emotional connotation of music, and the acceptance of users for duplicate recommendations. In the second part, we present our visions for future directions in MRS research (Section 3). More precisely, we elaborate on the topics of *psychologically-inspired music recommendation* (considering human personality and emotion), *situation-aware music recommendation*, and *culture-aware music recommendation*. We conclude this article with a summary and identification of possible starting points for the interested researcher to face the discussed challenges (Section 4).

This research was supported in part by the Austrian Science Fund (FWF): P25655, and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

Conference'17, Washington, DC, USA
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

The composition of the authors allows to take academic as well as industrial perspectives, which are both reflected in this article. Furthermore, we would like to highlight that particularly the ideas presented as *Challenge 2: Automatic playlist continuation* in Section 2 play an important role in the task definition, organization, and execution of the ACM Recommender Systems Challenge 2018[1] which focuses on this use case. This article may therefore also serve as an entry point for potential participants in this challenge.

2 GRAND CHALLENGES

In the following, we identify and detail a selection of the grand challenges, which we believe the research field of music recommender systems is currently facing, i.e., overcoming the cold start problem, automatic playlist continuation, and properly evaluating music recommender systems. We review the state of the art of the respective tasks and its current limitations.

2.1 Particularities of music recommendation

Before we start digging deeper into these challenges, we would first like to highlight the major aspects that make music recommendation a particular challenge and distinguishes it from recommending other items, such as movies, books, or products. These aspects have been adopted from a tutorial on music recommender systems [120], co-presented by one of the authors at the ACM Recommender Systems 2017 conference.[2]

Duration of items: While in traditional movie recommendation the items of interest have a typical duration of 90 minutes or more, the duration of music items usually ranges between 3 and 5 minutes (except for classical music). Because to this, items may be considered more disposable.

Magnitude of items: The size of common commercial music catalogs is in the range of tens of millions music pieces while movie streaming services have to deal with much smaller catalog sizes, typically thousands up to tens of thousands of movies and series. Scalability is therefore a much more important issue in music recommendation.

Sequential consumption: Unlike movies, music pieces are most frequently consumed sequentially, more than one at a time, i.e., in a listening session. This yields a number of challenges for a MRS, which relate to identifying the right arrangement of items in a recommendation list.

Repeated recommendations: Recommending the same music piece again, at a later point in time, may be appreciated by the user of a MRS, in contrast to a movie or product recommender, where repeated recommendations are usually not preferred.

Consumption behavior: Music is often consumed passively, in the background. While this is not a problem per se, it can affect preference elicitation. In particular when using implicit feedback to infer listener preferences, the fact that a listener is not paying attention to the music (therefore, e.g., not skipping a song) might be wrongly interpreted as a positive signal.

Listening intent and context: Listeners' intents for consuming or sharing a music piece are manifold and should be taken into account when building a MRS. For instance, a listener will likely create a different playlist when preparing for a romantic dinner than when warming-up with friends to go out at a Friday night. This also highlights the importance of the social component of music listening. Furthermore, a lot of listeners strongly identify with their liked artists. In this vein, music is also often used for self-expression. Another important and frequent intent is regulating the listener's mood, which is discussed below.
Similar to intent, the listening context may strongly influence the listeners' preferences. Among others, context may relate to location, for instance, listening at the workplace, when commuting, or relaxing at home. It might also relate to the use of different listening devices, e.g., earplugs on a smartphone vs. hi-fi stereo at home, just to give a few examples. The importance of considering such intent and context factors in MRS research is acknowledged by discussing situation-aware MRS as a trending research direction, cf. Section 3.2.

Emotional connotation: Music is known to evoke very strong emotions, and one of the most frequent reasons to listen to music is indeed mood regulation [93]. At the same time, the emotion of the listener is usually neglected in current MRS. This is the reason why we selected emotion-aware MRS as one of the main future directions in MRS research, cf. Section 3.1.

2.2 Challenge 1: Cold start problem

Problem definition: One of the major problems of recommender systems in general [43, 112], and music recommender systems in particular [73, 89] is the *cold start* problem, i.e., when a new user registers to the system or a new item is added to the catalog and the system does not have sufficient data associated with these items/users. In such a case, the system cannot properly recommend existing items to a new user (*new user* problem) or recommend a new item to the existing users (*new item* problem) [3, 43, 73, 123].

Another sub-problem of cold start is the *sparsity* problem which refers to the fact that the number of given ratings is much lower than the number of possible ratings, which is particularly likely when the number of users and items is large. The inverse of the ratio between given and possible ratings is called sparsity. High sparsity translates into low rating *coverage*, since most users tend to rate only a tiny fraction of items. The effect is that recommendations often become unreliable [73]. Typical values of sparsity are quite close to 100% in most real-world recommender systems. In the music domain, this is a particularly substantial problem. Dror et al. [34], for instance, analyzed the Yahoo! Music dataset, which as of time of writing represents the largest music recommendation dataset. They report a sparsity of 99.96%. For comparison, the Netflix dataset of movies has a sparsity of "only" 98.82%.

State of the art:

A number of approaches have already been proposed to tackle the cold start problem in the music recommendation domain, foremost content-based approaches, hybridization, cross-domain recommendation, and active learning.

Content-based recommendation (CB) algorithms do not require the rating of other users. Therefore, as long as some pieces of information about the user's own preferences are available, such techniques can be used in cold start scenarios. Furthermore, in the

[1] https://recsys.acm.org/recsys18/challenge
[2] http://www.cp.jku.at/tutorials/mrs_recsys_2017

most severe case, when a new item is added to the catalog, content-based methods enable recommendations, because they can extract features from the new item and use them to make recommendations. It is noteworthy that while collaborative filtering (CF) systems have cold start problems both for new users and new items, content-based systems have only cold start problems for new users [5].
As for the new item problem, a standard approach is to extract a number of features that define the acoustic properties of the audio signal and use content-based learning of the user interest (user profile learning) in order to effect recommendations. This is advantageous not only to address the new item problem but also because an accurate feature representation can be highly predicative of users' tastes and interests which can be leveraged in the subsequent information filtering stage [5]. Such feature extraction from audio signals can be done in two main manners: (1) by extracting a feature vector from each item individually, independent of other items or (2) by considering the cross-relation between items in the training dataset. The difference is that in (1) the same process is performed in the training and testing phases of the system, and the extracted feature vectors can be used off-the-shelf in the subsequent processing stage, for example they can be used to compute similarities between items in a one-to-one fashion at testing time. In contrast, in (2) first a model is built from all features extracted in the training phase, whose main role is to map the features into a new (acoustic) space in which the similarities between items are better represented and exploited. An example of approach (1) is the block-level feature framework [126, 127], which creates a feature vector of about 10,000 dimensions, independently for each song in the given music collection. This vector describes aspects such as spectral patterns, recurring beats, and correlations between frequency bands. An example of strategy (2) is to create a low-dimensional i-vector representation from the Mel-frequency cepstral coefficients (MFCC), which model musical timbre to some extent [39]. To this end, a universal background model is created from the MFCC vectors of the whole music collection, using a Gaussian mixture model (GMM). Performing factor analysis on a representation of the GMM eventually yields i-vectors.

In scenarios where some form of semantic labels, e.g., genres or musical instruments, are available, it is possible to build models that learn the intermediate mapping between low-level audio features and semantic representations using machine learning techniques, and subsequently use the learned models for prediction. A good point of reference for such *semantic-inferred* approaches can be found in [13, 24].

An alternative technique to tackle the new item problem is *hybridization*. A review of different hybrid and ensemble recommender systems can be found in [6, 18]. In [33] the authors propose a music recommender system which combines an acoustic CB and an item-based CF recommmender. For the content-based component, it computes acoustic features including spectral properties, timbre, rhythm, and pitch. The content-based component then assists the collaborative filtering recommender in tackling the cold start problem since the features of the former are automatically derived via audio content analysis.
The solution proposed in [146] is a hybrid recommender system that combines CF and acoustic CB strategies also by feature hybridization. However, in this work the feature-level hybridization is not done in the original feature domain. Instead, a set of latent variables referred to as *conceptual genre* are introduced, whose role is to provide a common shared feature space for the two recommenders and enable hybridization. The weights associated with the latent variables reflect the musical taste of the target user and are learned during the training stage.
In [128] the authors propose a hybrid recommender system incorporating item–item CF and acoustic CB based on similarity metric learning. The proposed metric learning is an optimization model that aims to learn the weights associated with the audio content features (when combined in a linear fashion) so that a degree of consistency between CF-based similarity and the acoustic CB similarity measure is established. The optimization problem can be solved using quadratic programming techniques.

Another solution to cold start are *cross-domain recommendation* techniques, which aim at improving recommendations in one domain (here music) by making use of information about the user preferences in an auxiliary domain [20, 46]. Hence, the knowledge of the preferences of the user is transferred from an auxiliary domain to the music domain, resulting in a more complete and accurate user model. Similarly, it is also possible to integrate additional pieces of information about the (new) users, which are not directly related to music, such as their personality, in order to improve the estimation of the user's music preferences. Several studies conducted on user personality characteristics support the conjecture that it may be useful to exploit this information in music recommender systems [48, 52, 63, 98, 108]. For a more detailed literature review of cross-domain recommendation, we refer to [21, 47, 76].

In addition to the aforementioned approaches, *active learning* has shown promising results in dealing with the cold start problem. Active learning addresses this problem at its origin by identifying and eliciting (high quality) data that can represent the preferences of users better than by what they provide themselves [43, 112]. Such a system therefore interactively demands specific user feedback to maximize the improvement of system performance.

Limitations: The state-of-the-art approaches discussed above are restricted by certain limitations. When using *content-based filtering*, for instance, almost all existing approaches rely on a number of predefined audio features that have been used over and over again, including spectral features, MFCCs, and a great number of derivatives [80]. However, doing so assumes that (all) these features are predictive of the user's music taste, while in practice it has been shown that the acoustic properties that are important for the perception of music are highly subjective [100]. Furthermore, listeners' different tastes and amount of interests in different pieces of music influence perception of item similarity [117]. This subjectiveness demands for CB recommenders that incorporate personalization in their mathematical model. For example, in [44] the authors propose a hybrid (CB+CF) recommender model, namely regression-based latent factor models (RLFM). In [4] the authors propose a user-specific feature-based similarity model (UFSM), which defines a similarity function for each user, leading to a high degree of personalization. Although not designed specifically for the music domain, the authors of [4] provide an interesting literature review of similar user-specific models.

While *hybridization* can therefore alleviate the cold start problem to a certain extent, as seen in the examples above, respective approaches are often complex, computationally expensive, and lack transparency [19]. In particular, results of hybrids employing latent factor models are typically hard to understand for humans.

A major problem with *cross-domain recommender systems* is their need for data that connects two or more target domains, e.g., books, movies, and music [21]. In order for such approaches to work properly, items, users, or both therefore need to overlap to a certain degree [27]. In the absence of such overlap, relationships between the domains must be established otherwise, e.g., by inferring semantic relationships between items in different domains or assuming similar rating patterns of users in the involved domains. However, whether respective approaches are capable of transferring knowledge between domains is disputed [26]. A related issue in cross-domain recommendation is that there is a lack of established datasets with clear definition of domains and recommendation scenarios [76]. Because of this, the majority of existing work on cross-domain RS use some type of conventional recommendation dataset transformation to suit it for their need.

Finally, also *active learning* techniques suffer from a number of issues. First of all, the typical active learning techniques propose to the users the items with the highest predicted ratings in order to elicit the true ratings. This indeed is a default strategy in recommender systems as users tend to rate what have been recommended to them. Moreover, users typically browse and rate interesting items which they would like. However, it has been shown that doing so creates a strong bias in the dataset and expands it disproportionately with high ratings. This in turn may substantially influence the prediction algorithm and decrease the recommendation accuracy [42]. Moreover, not all the active learning strategies are necessarily personalized. The users differ very much in the amount of information they have about the items, their preferences, and the way they make decisions. Hence, it is clearly inefficient to request all the users to rate the same set of items, because many users may have a very limited knowledge, ignore many items, and not properly provide ratings for these items. Properly designed active learning techniques should take this into account and propose different items to different users to rate. This can be very beneficial and increase the chance of acquiring ratings of higher quality [40].

2.3 Challenge 2: Automatic playlist continuation

Problem definition: In its most generic definition, a playlist is simply a sequence of tracks intended to be listened to together. The task of automatic playlist generation (APG) then refers to the automated creation of these sequences of tracks.

Considered a variation of APG, the task of *automatic playlist continuation* (APC) consists of adding one or more tracks to a playlist in a way that fits the same target characteristics of the original playlist. This has benefits in both the listening and creation of playlists: users can enjoy listening to continuous sessions beyond the end of a finite-length playlist, while also finding it easier to create longer, more compelling playlists without needing to have extensive musical familiarity.

A large part of the APC task is to accurately infer the intended purpose of a given playlist. This is challenging not only because of the broad range of these intended purposes (when they even exist), but also because of the diversity in the underlying features or characteristics that might be needed to infer those purposes.

Related to Challenge 1, an extreme cold start scenario for this task is where a playlist is created with some metadata (a title, for example), but no song has been added to the playlist. This problem can be cast as an *ad-hoc information retrieval task*, where the task is to rank songs in response to a user-provided metadata query.

The APC task can also potentially benefit from user profiling, e.g., making use of previous playlists and the long-term listening history of the user. We call this *personalized playlist continuation.*

According to a study carried out in 2016 by the Music Business Association[3] as part of their Music Biz Consumer Insights program,[4] playlists accounted for 31% of music listening time among listeners in the USA, more than albums (22%), but less than single tracks (46%). Other studies, conducted by MIDiA,[5] show that 55% of streaming music service subscribers create music playlists, with some streaming services such as Spotify currently hosting over 2 billion playlists.[6] Studies like these suggest a growing importance of playlists as a mode of music consumption, and as such, the study of APG and APC has never been more relevant.

State of the art: APG has been studied ever since digital multimedia transmission made huge catalogs of music available to users. Bonnin and Jannach provide a comprehensive survey of this field in [15]. In it, the authors frame the APG task as the creation of a sequence of tracks that fulfill some "target characteristics" of a playlist, given some "background knowledge" of the characteristics of the catalog of tracks from which the playlist tracks are drawn. Existing APG systems tackle both of these problems in many different ways.

In early approaches [8, 9, 101] the target characteristics of the playlist are specified as multiple explicit constraints, which include musical attributes or metadata such as artist, tempo, and style. In others, the target characteristics are a single seed track [92] or a start and an end track [8, 22, 53]. Other approaches create a circular playlist that comprises all tracks in a given music collection, in such a way that consecutive songs are as similar as possible [79, 107]. In other works, playlists are created based on the context of the listener, either as single source [116] or in combination with content-based similarity [23, 110].

A common approach to build the background knowledge of the music catalog for playlist generation is using machine learning techniques to extract that knowledge from manually-curated playlists. The assumption here is that curators of these playlists are encoding rich latent information about which tracks go together to create a satisfying listening experience for an intended purpose. Some proposed APG and APC systems are trained on playlists from such sources as online radio stations [22, 94], online playlist websites [95, 139], and music streaming services [106]. In the study by Pichl et al. [106], the names of playlists on Spotify were analyzed

[3] https://musicbiz.org/news/playlists-overtake-albums-listenership-says-loop-study
[4] https://musicbiz.org/resources/tools/music-biz-consumer-insights/consumer-insights-portal
[5] https://www.midiaresearch.com/blog/announcing-midias-state-of-the-streaming-nation-2-report
[6] https://press.spotify.com/us/about

to create contextual clusters, which were then used to improve recommendations.

Limitations: While some work on automated playlist continuation highlights the special characteristics of playlists, i.e., their *sequential order*, it is not well understood to which extent and in which cases taking into account the order of tracks in playlists helps create better models for recommendation. For instance, in [139] Vall et al. recently demonstrated on two datasets of hand-curated playlists that the song order seems to be negligible for accurate playlist continuation when a lot of popular songs are present. On the other hand, the authors argue that order does matter when creating playlists with tracks from the long tail. Another study by McFee and Lanckriet [95] also suggests that transition effects play an important role in modeling playlist continuity. In another recent user study [135] conducted by Tintarev et al., the authors found that many participants did not care about the order of tracks in recommended playlists, sometimes they did not even notice that there is a particular order. However, this study was restricted to 20 participants who used the Discover Weekly service of Spotify.[7]

Another challenge for APC is evaluation: in other words, how to assess the quality of a playlist. Evaluation in general is discussed in more detail in the next section, but there are specific questions around evaluation of playlists in particular that should be pointed out here. As Bonnin and Jannach [15] put it, the ultimate criterion for this is *user satisfaction*, but that is not easy to measure. In [96], McFee and Lanckriet categorize the main approaches to APG evaluation as human evaluation, semantic cohesion, and sequence prediction. Human evaluation comes closest to measuring user satisfaction directly, but suffers from problems of scale and reproducibility. Semantic cohesion as a quality metric is easily measurable and reproducible, but assumes that users prefer playlists where tracks are similar along a particular semantic dimension, which may not always be true, see for instance the studies carried out by Slaney and White [131] and by Lee [88]. Sequence prediction casts APC as an information retrieval task, but in the domain of music, an inaccurate prediction need not be a bad recommendation, and this again leads to a potential disconnect between this metric and the ultimate criterion of user satisfaction.

Investigating which factors are potentially important for a positive user perception of a playlist, Lee conducted a qualitative user study [88], investigating playlists that had been automatically created based on content-based *similarity*. They made several interesting observations. A concern frequently raised by participants was that of consecutive songs being too similar, and a general lack of *variety*. However, different people had different interpretations of variety, e.g., variety in genres or styles vs. different artists in the playlist. Similarly, different criteria were mentioned when listeners judged the coherence of songs in a playlist, including lyrical content, tempo, and mood. When creating playlists, participants mentioned that similar lyrics, a common theme (e.g., music to listen to in the train), story (e.g., music for the Independence Day), or era (e.g., rock music from the 1980s) are important and that tracks not complying negatively effect the flow of the playlist. These aspects can be extended by responses of participants in a study conducted by Cunningham et al. [29], who further identified the following categories of playlists: same artist, genre, style, or orchestration, playlists for a certain event or activity (e.g., party or holiday), romance (e.g., love songs or breakup songs), playlists intended to send a message to their recipient (e.g., protest songs), and challenges or puzzles (e.g., cover songs liked more than the original or songs whose title contains a question mark).

Lee also found that *personal preferences* play a major role. In fact, already a single song, which is very much liked or hated by a listener, can have a strong influence on how they judge the entire playlist [88], in particular if it is a highly disliked song [31]. Furthermore, a good *mix of familiar and unknown songs* was often mentioned as an important requirement for a good playlist. Supporting the discovery of interesting new songs, still contextualized by familiar ones, increases the *serendipity* [119, 149] of a playlist. Finally, participants also reported that their familiarity with a playlist's genre or theme influenced their judgment of its quality. In general, listeners were more picky about playlists whose tracks they were familiar with or they liked a lot.

Supported by the studies summarized above, we argue that the question of what makes a great playlist is highly subjective and further depends on the intent of the creator or listener. Important criteria when creating or judging a playlist include track similarity/coherence, variety/diversity, but also the user's personal preferences and familiarity with the tracks, as well as the intention of the playlist creator. Unfortunately, current automatic approaches to playlist continuation are agnostic of the underlying psychological and sociological factors that influence the decision of which songs users choose to include in a playlist. Since knowing about such factors is vital to understand the intent of the playlist creator, we believe that algorithmic methods for automatic playlist continuation need to holistically learn such aspects from manually created playlists and integrate respective intent models. However, we are aware that in today's era where billions of playlists are shared by users of online streaming services,[8] a large-scale analysis of psychological and sociological background factors is impossible. Nevertheless, in the absence of explicit information about user intent, a possible starting point to create intent models might be the metadata associated with user-generated playlists, such as title or description. To foster this kind of research, the playlists provided in the dataset for the ACM Recommender Systems Challenge 2018 will include playlist titles.[9]

2.4 Challenge 3: Evaluating music recommender systems

Problem definition: Having its roots in machine learning (cf. rating prediction) and information retrieval (cf. "retrieving" items based on implicit "queries" given by user preferences), the field of recommender systems originally adopted evaluation metrics from these neighboring fields. In fact, accuracy and related quantitative measures, such as precision, recall, or error measures (between predicted and true ratings), are still the most commonly employed criteria to judge the recommendation quality of a recommender system [10, 56]. In addition, novel measures that are

[7] https://www.spotify.com/discoverweekly

[8] https://press.spotify.com/us/about

[9] https://recsys.acm.org/recsys18/challenge

tailored to the recommendation problem and often take a user-centric perspective have emerged in recent years. These so-called beyond-accuracy measures [72] address the particularities of recommender systems and gauge, for instance, the utility, novelty, or serendipity of an item for a user. However, a major problem with these kinds of measures is that they integrate factors that are hard to describe mathematically, for instance, the aspect of surprise in case of serendipity measures. For this reason, there sometimes exist a variety of different definitions to quantify the same beyond-accuracy aspect.

State of the art: The following performance measures are the ones most frequently reported when evaluating recommender systems. They can be roughly categorized into accuracy-related measures, such as prediction error (e.g., MAE and RMSE) or standard IR measures (e.g., precision and recall), and beyond-accuracy measures, such as diversity, novelty, and serendipity. Furthermore, while some of the metrics quantify the ability of recommender systems to find good items, e.g., precision, MAE, or RMSE, others consider the ranking of items and therefore assess the system's ability to position good recommendations at the top of the recommendation list, e.g., MAP, NDCG, or MPR.

Mean absolute error (MAE) is one of the most common metrics for evaluating the prediction power of recommender algorithms. It computes the average absolute deviation between the predicted ratings and the actual ratings provided by users [58]. Indeed, MAE indicates how close the rating predictions generated by an MRS are to the real user ratings. MAE is computed as follows:

$$MAE = \frac{1}{|T|} \sum_{r_{u,i} \in T} |r_{u,i} - \hat{r}_{u,i}| \tag{1}$$

where $r_{u,i}$ and $\hat{r}_{u,i}$ respectively denote the actual and the predicted ratings of item i for user u. MAE sums over the absolute prediction errors for all ratings in a test set T.

Root mean square error (RMSE) is another similar metric that is computed as:

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{r_{u,i} \in T} (r_{u,i} - \hat{r}_{u,i})^2} \tag{2}$$

It is an extension to MAE in that the error term is squared, which penalizes larger differences between predicted and true ratings more than smaller ones. This is motivated by the assumption that, for instance, a rating prediction of 1 when the true rating is 4 is much more severe than a prediction of 3 for the same item.

Precision at top K recommendations (P@K) is a common metric that measures the accuracy of the system in commanding relevant items. In order to compute the precision, for each user, the top K recommended items whose ratings also appear in test set T are considered. This metric was originally designed for binary relevance judgments. Therefore, in case of availability of relevance information at different levels, such as a five point Likert scale, the labels should be binarized, e.g., considering the ratings greater than or equal to 4 (out of 5) as relevant. Precision@K is computed as follows:

$$P@K = \frac{1}{|U|} \sum_{u \in U} \frac{|L_u \cap \hat{L}_u|}{|\hat{L}_u|} \tag{3}$$

where L_u is a set of relevant items of user u in the test set T and $\hat{L}_u$ denotes the recommended set containing the K items in T with the highest predicted ratings for the user u from the set of all users U.

Mean average precision (MAP) is a metric that computes the overall precision of a recommender system based on precision at different recall levels [90]. It is computed as the arithmetic mean of the average precision (AP) over the entire set of users in the test set, where AP is defined as follows:

$$AP = \frac{1}{\min(M, N)} \sum_{k=1}^{N} P@k \cdot rel(k) \tag{4}$$

where $rel(k)$ is an indicator signaling if the k^{th} recommended item is relevant, i.e. $rel(k) = 1$, or not, i.e. $rel(k) = 0$; M is the number of relevant items and N is the number of recommended items in the top N recommendation list. Note that AP implicitly incorporates recall, because it considers relevant items not in the recommendation list.

Recall at top K recommendations (R@K) is presented here for the sake of completeness, even though it is not a crucial measure from a consumer's perspective. Indeed, the listener is typically not interested in being recommended all or a large number of relevant items, rather in having good recommendations at the top of the recommendation list. R@K is defined as:

$$R@K = \frac{1}{|U|} \sum_{u \in U} \frac{|L_u \cap \hat{L}_u|}{|L_u|} \tag{5}$$

where L_u is a set of relevant items of user u in the test set T and $\hat{L}_u$ denotes the recommended set containing the K items in T with the highest predicted ratings for the user u from the set of all users U.

Normalized discounted cumulative gain (NDCG) measures the ranking quality of the recommendations. This metric has originally been proposed to evaluate effectiveness of information retrieval systems [69]. It is nowadays also frequently used for evaluating music recommender systems [91, 104, 143]. Assuming that the recommendations for user u are sorted according to the predicted rating values in descending order. DCG_u is defined as follows:

$$DCG_u = \sum_{i=1}^{N} \frac{r_{u,i}}{log_2(i+1)} \tag{6}$$

where $r_{u,i}$ is the true rating (as found in test set T) for the item ranked in position i for user u, and N is the length of the recommendation list. Since the rating distribution depends on the users' behavior, the DCG values for different users are not directly comparable. Therefore, the cumulative gain for each user should be normalized. This is done by computing the ideal DCG for user u, denoted as $IDCG_u$, which is the DCG_u value for the best possible ranking, obtained by ordering the items by true ratings in descending order. Normalized discounted cumulative gain for user u is then calculated as:

$$NDCG_u = \frac{DCG_u}{IDCG_u} \tag{7}$$

Finally, the overall normalized discounted cumulative gain *NDCG* is computed by averaging $NDCG_u$ over the entire set of users.

In the following, we present common quantitative evaluation metrics, which have been particularly designed or adopted to assess

recommender systems performance, even though some of them have their origin in information retrieval and machine learning.

Half life utility (HLU) measures the utility of a recommendation list for a user with the assumption that the likelihood of viewing/choosing a recommended item by the user exponentially decays with the item's position in the ranking [16, 102]. Formally written, HLU for user u is defined as:

$$HLU_u = \sum_{i=1}^{N} \frac{\max(r_{ui} - d, 0)}{2^{(rank_{ui}-1)/(h-1)}} \tag{8}$$

where r_{ui} and $rank_{ui}$ denote the rating and the rank of item i for user u, respectively, in the recommendation list of length N; d represents a default rating (e.g., average rating) and h is the half-time, calculating as the rank of a music item in the list, such that the user can eventually listen to it with a 50% chance. HLU_u can be further normalized by the maximum utility (similar to NDCG), and the final HLU is the average over the half-time utilities obtained for all users in the test set. A larger HLU may correspond to a superior recommendation performance.

Mean percentile rank (MPR) estimates the users' satisfaction with items in the recommendation list, and is computed as the average of the percentile rank for each test item within the ranked list of recommended items for each user [66]. The percentile rank of an item is the percentage of items whose position in the recommendation list is equal to or lower than the position of the item itself. Formally, the percentile rank PR_u for user u is defined as:

$$PR_u = \frac{\sum_{r_{u,i} \in T} r_{u,i} \cdot rank_{u,i}}{\sum_{r_{u,i} \in T} r_{u,i}} \tag{9}$$

where $r_{u,i}$ is the true rating (as found in test set T) for item i rated by user u and $rank_{u,i}$ is the percentile rank of item i within the ordered list of recommendations for user u. MPR is then the arithmetic mean of the individual PR_u values over all users. A randomly ordered recommendation list has an expected MPR value of 50%. A smaller MPR value is therefore assumed to correspond to a superior recommendation performance.

Spread is a metric of how well the recommender algorithm can spread its attention across a larger set of items [78]. In more detail, spread is the entropy of the distribution of the items recommended to the users in the test set. It is formally defined as:

$$spread = -\sum_{i \in I} P(i) \log P(i) \tag{10}$$

where I represents the entirety of items in the dataset and $P(i) = count(i)/\sum_{i' \in I} count(i')$, such that $count(i)$ denotes the total number of times that a given item i showed up in the recommendation lists. It may be infeasible to expect an algorithm to achieve the perfect spread (i.e., recommending each item an equal number of times) without avoiding irrelevant recommendations or unfulfillable rating requests. Accordingly, moderate spread values are usually preferable.

Coverage of a recommender system is defined as the proportion of items over which the system is capable of generating recommendations [58]:

$$coverage = \frac{|\hat{T}|}{|T|} \tag{11}$$

where $|T|$ is the size of the test set and $|\hat{T}|$ is the number of ratings in T for which the system can predict a value. This is particularly important in cold start situations, when recommender systems are not able to accurately predict the ratings of new users or new items, and hence obtain low coverage. Recommender systems with lower coverage are therefore limited in the number of items they can recommend. A simple remedy to improve low coverage is to implement some default recommendation strategy for an unknown user–item entry. For example, we can consider the average rating of users for an item as an estimate of its rating. This may come at the price of accuracy and therefore the trade-off between coverage and accuracy needs to be considered in the evaluation process [7].

Novelty measures the ability of a recommender system to recommend new items that the user did not know about before [1]. A recommendation list may be accurate, but if it contains a lot of items that are not novel to a user, it is not necessarily a useful list [149].

While novelty should be defined on an individual user level, considering the actual freshness of the recommended items, it is common to use the self-information of the recommended items relative to their global popularity:

$$novelty = \frac{1}{|U|} \sum_{u \in U} \sum_{i \in L_u} \frac{\log_2 pop_i}{N} \tag{12}$$

where pop_i is the popularity of item i measured as percentage of users who rated i, L_u is the recommendation list of the top N recommendations for user u [149, 151]. The above definition assumes that the likelihood of the user selecting a previously unknown item is proportional to its global popularity and is used as an approximation of novelty. In order to obtain more accurate information about novelty or freshness, explicit user feedback is needed, in particular since the user might have listened to an item through other channels before.

It is often assumed that the users prefer recommendation lists with more novel items. However, if the presented items are too novel, then the user is unlikely to have any knowledge of them, nor to be able to understand or rate them. Therefore, moderate values indicate better performances [78].

Serendipity aims at evaluating MRS based on the *relevant and surprising* recommendations. While the need for serendipity is commonly agreed upon [59], the question of how to measure the degree of serendipity for a recommendation list is controversial. This particularly holds for the question of whether the factor of surprise implies that items must be novel to the user [72]. On a general level, serendipity of a recommendation list L_u provided to a user u can be defined as:

$$serendipity(L_u) = \frac{\left| L_u^{unexp} \cap L_u^{useful} \right|}{|L_u|} \tag{13}$$

where L_u^{unexp} and L_u^{useful} denote subsets of L that contain, respectively, recommendations unexpected to and useful for the user. The

usefulness of an item is commonly assessed by explicitly asking users or taking user ratings as proxy [72]. The unexpectedness of an item is typically quantified by some measure of distance from expected items — those similar to the items already rated by the user. In the context of MRS, Zhang et al. [149] propose an "unserendipity" measure that is defined as the average similarity between the items in the user's listening history and the new recommendations. Similarity between two items in this case is calculated by an adapted cosine measure that integrates co-liking information, i.e., number of users who like both items. It is assumed that lower values correspond to more surprising recommendations, since lower values indicate that recommendations deviate from the userfis traditional behavior [149].

Diversity is another important beyond-accuracy measure as already discussed in the limitations part of Challenge 1. It gauges the extent to which recommended items are different from each other, where difference can relate to various aspects, e.g., musical style, artist, lyrics, or instrumentation, just to name a few. Similar to serendipity, diversity can be defined in several ways. One of the most common is to compute pairwise distance between all items in the recommendation set, either averaged [152] or summed [132]. In the former case, the diversity of a recommendation list L is calculated as follows:

$$diversity(L) = \frac{\sum_{i \in L} \sum_{j \in L \setminus i} dist_{i,j}}{|L| \cdot (|L| - 1)} \tag{14}$$

where $dist_{i,j}$ is the some distance function defined between items i and j. Common choices are inverse cosine similarity [111], inverse Pearson correlation [141], or Hamming distance [75].

Limitations: As of today, the vast majority of evaluation approaches in recommender systems research focuses on quantitative measures, either accuracy-like or beyond-accuracy, which are often computed in offline studies. While doing so has the advantage of facilitating the reproducibility of evaluation results, these approaches typically fall short of grasping some of the most important user requirements that relate to user acceptance or satisfaction, among others.

Despite acknowledging the need for more user-centric evaluation strategies [117], the factor human, user, or, in the case of MRS, listener is still way too often neglected or not properly addressed. For instance, while there exist quantitative measures for serendipity and diversity, as discussed above, *perceived* serendipity and diversity can be highly different from the measured ones [140]. Even beyond-accuracy measures can therefore not fully capture the real *user satisfaction* with a recommender system.

Addressing both objective and subjective evaluation criteria, Knijnenburg et al. [81] propose a holistic framework for user-centric evaluation of recommender systems. Figure 1 provides an overview of the components. The objective system aspects (OSA) are considered unbiased factors of the RS, including aspects of the user interface, computing time of the algorithm, or number of items shown to the user. They are typically easy to specify or compute. The OSA influence the subjective system aspects (SSA), which are caused by momentary, primary evaluative feelings while interacting with the system [57]. This results in a different perception of the system by different users. SSA are therefore highly individual aspects and typically assessed by user questionnaires. Examples of SSA include general appeal of the system, usability, and perceived recommendation diversity or novelty. The aspect of experience (EXP) describes the user's attitude towards the system and is commonly also investigated by questionnaires. It addresses the user's perception of the interaction with the system. The experience is highly influenced by the other components, which means changing any of the other components likely results in a change of EXP aspects. Experience can be broken down into the evaluation of the system, the decision process, and the final decisions made, i.e., the outcome. The interaction (INT) aspects describe the observable behavior of the user, such as time spent viewing an item, clicking or purchasing behavior. Therefore, they belong to the objective measures and are usually determined via logging by the system. Finally, Knijnenburg et al.'s framework mentions personal characteristics (PC) and situational characteristics (SC), which influence the user experience. PC include aspects that do not exist without the user, such as user demographics, knowledge, or perceived control, while SC include aspects of the interaction context, such as when and where the system is used, or situation-specific trust or privacy concerns. Knijnenburg et al. [81] also propose a questionnaire to asses the factors defined in their framework, for instance, perceived recommendation quality, perceived system effectiveness, perceived recommendation variety, choice satisfaction, intention to provide feedback, general trust in technology, and system-specific privacy concern.

While this framework is a generic one, tailoring it to MRS would allow for user-centric evaluation thereof. Especially the aspects of personal and situational characteristics should be adapted to the particularities of music listeners and listening situations, respectively, cf. Section 2.1. To this end, researchers in MRS should consider the aspects relevant for the perception and preference of music, and their implications on MRS, which have been identified in several studies, e.g. [30, 86, 87, 117, 118]. In addition to the general ones mentioned by Knijnenburg et al., of great importance in the music domain seem to be psychological factors, including affect and personality, social influence, musical training and experience, and physiological condition.

We believe that carefully and holistically evaluating MRS by means of accuracy and beyond-accuracy, objective and subjective measures, in offline and online experiments, would lead to a better understanding of the listeners' needs and requirements vis-à-vis MRS, and eventually a considerable improvement of current MRS.

3 FUTURE DIRECTIONS AND VISIONS

While the challenges identified in the previous section are already researched on intensely, in the following, we provide a more forward-looking analysis and discuss some MRS-related trending topics, which we assume influential for the next generation of MRS. All of them have in common that their aim is to create more personalized recommendations. More precisely, we first outline how psychological constructs such as personality and emotion could be integrated into MRS. Subsequently, we address situation-aware MRS and argue for the need of multifaceted user models that describe contextual

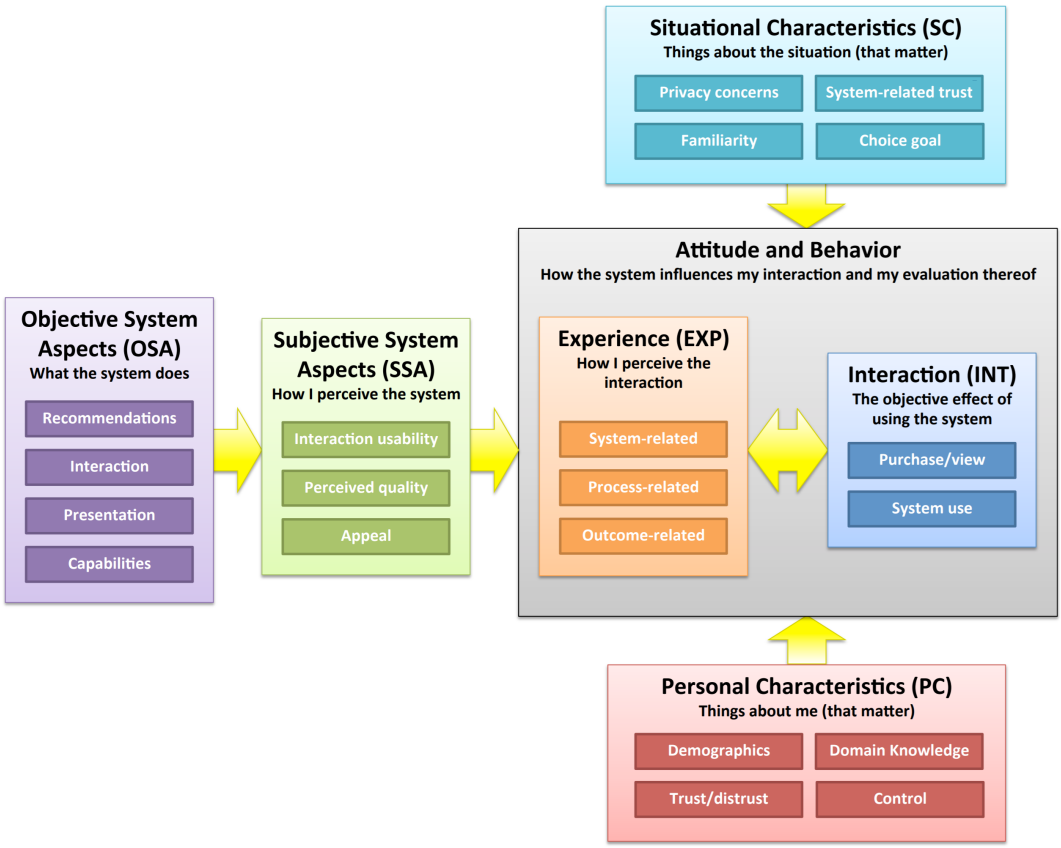

Figure 1: Evaluation framework of the user experience for recommender systems, according to [81].

and situational preferences. To round off, we discuss the influence of users' cultural background on recommendation preferences, which needs to be considered when building culture-aware MRS.

3.1 Psychologically-inspired music recommendation

Personality and emotion are important psychological constructs. While personality characteristics of humans are a predictable and stable measure that shapes human behaviors, emotions are short-term affective responses to a particular stimulus [137]. Both have been shown to influence music tastes [50, 114, 118] and user requirements for MRS [48, 52]. However, in the context of (music) recommender systems, personality and emotion do not play a major role yet. Given the strong evidence that both influence listening preferences [108, 118] and the recent emergence of approaches to accurately predict them from user-generated data [83, 129], we believe that psychologically-inspired MRS is an upcoming area.

Personality: In psychology research, personality is often defined as a "consistent behavior pattern and interpersonal processes originating within the individual" [17]. This definition accounts for the individual differences in people's emotional, interpersonal, experiential, attitudinal, and motivational styles [71]. Several prior works have studied the relation of decision making and personality factors. In [108], as an example, it has been shown that personality can influence the human decision making process as well as the tastes and interests. Due to this direct relation, people with similar personality factors are very likely to share similar interests and tastes.

Earlier studies conducted on the user personality characteristics support the potential benefits that personality information could have in recommender systems [41, 62, 64, 136, 138]. As a known example, psychological studies [108] have shown that extravert people are likely to prefer the upbeat and conventional music. Accordingly, a personality-based MRS could use this information to better predict which songs are more likely than others to please extravert people [63]. Another example of potential usage is to exploit personality information in order to compute similarity among users and hence identify the like-minded users [136]. This similarity information could then be integrated into a neighborhood-based collaborative filtering approach.

In order to use personality information in a recommender system, the system first has to elicit this information from the users, which can be done either explicitly or implicitly. In the former case, the system can ask the user to complete a personality questionnaire using one of the personality evaluation inventories, e.g., the ten item personality inventory [55] or the big five inventory [70]. In the latter case, the system can learn the personality by tracking and observing users' behavioral patterns [84, 129]. Not too surprisingly,

it has shown that systems that explicitly elicit personality characteristics achieve superior recommendation outcomes, e.g., in terms of user satisfaction, ease of use, and prediction accuracy [35]. On the downside, however, many users are not willing to fill in long questionnaires before being able to use the RS. A way to alleviate this problem is to ask users only the most informative questions of a personality instrument [122]. Which questions are most informative, though, first needs to be determined based on existing user data and is dependent on the recommendation domain. Other studies showed that users are to some extent willing to provide further information in return for a better quality of recommendations [134].

Personality information can be used in various ways, particularly, to generate recommendations when traditional rating or consumption data is missing. Otherwise, the personality traits can be seen as an additional feature that extends the user profile, that can be used mainly to identify similar users in neighborhood-based recommender systems or directly fed into extended matrix factorization models [46].

Emotion: The emotional state of the MRS user is an important factor in identifying his or her short-time musical preferences, in particularly since emotion regulation is known to be one main reason why people listen to music [93]. Indeed, a music piece can be seen as an emotion-laden content and in turn can be described by emotions. Musical content contains various elements that can effect the emotional state of a person, such as rhythm, key, tempo, melody, harmony, and lyrics. For instance, a musical piece that is in major key is typically perceived brighter and happier than those in minor key, or a piece in rapid tempo is perceived more exciting or more tense than slow tempo ones [85].

Several studies have already shown that listeners' emotions have a strong impact on their musical preferences [73]. As an example, people may listen to completely different musical genres or styles when they are sad in comparison to when they are happy. Indeed, prior research on music psychology discovered that people may choose the type of music which moderates their emotional condition [82]. More recent findings show that music can be mainly chosen so as to augment the emotional situation perceived by the listener [99].

Similar to personality traits, the emotional state of a user can be elicited explicitly or implicitly. In the former case, the user is typically presented one of the various *categorical models* (emotions are described by distinct emotion words such as happiness, sadness, anger, or fear) [61, 148] or *dimensional models* (emotions are described by scores with respect to two or three dimensions, e.g., valence and arousal) [113]. For a more detailed elaboration on emotion models in the context of music, we refer to [118, 144]. The implicit acquisition of emotional states can be effected, for instance, by analyzing user-generated text [32], speech [45], or facial expressions in video [38].

Since music can be viewed as an emotionally laden content, as it is capable of evoking intense emotions in a listener, it can also be annotated with emotional labels [68, 145, 148]. Doing so automatically is a task referred to as music emotion recognition (MER) and is discussed in detail, for instance, in [77, 144]. While such automatic emotion labeling of music items could be beneficial for MRS, MER has been shown to be a highly challenging task [77].

Nowadays, emotion-based recommender systems typically consider emotional scores as contextual factors that characterize the contextual situation that the user is experiencing. Hence, the recommender systems exploit emotions in order to pre-filter the preferences of users or post-filter the generated recommendations. Unfortunately, this neglects the psychological background, in particular on the subjective and complex interrelationships between expressed, perceived, and induced emotions [118], which is of special importance in the music domain as music is known to evoke stronger emotions than, for instance, products [120]. It has also been shown that personality influences in which emotional state which kind of emotionally laden music is preferred by listeners [50]. Therefore, even if automated MER approaches would be able to accurately predict the perceived or induced emotion of a given music piece, in the absence of deep psychological listener profiles, matching emotion annotations of items and listeners may not yield satisfying recommendations. We hence believe that the field of MRS should embrace psychological theories, elicit the respective user-specific traits, and integrate them into recommender systems, in order to build decent emotion-aware MRS.

3.2 Situation-aware music recommendation

Most of the existing music recommender systems make recommendations solely based on a set of user-specific and item-specific signals. However, in real-world scenarios, many other signals are available. These additional signals can be further used to improve the recommendation performance. A large subset of these additional signals includes *situational signals*. In more detail, the music taste of a user depends on the situation at the moment of recommendation. *Location* is an example of situational signals; for instance, the music taste of a user would differ in libraries and in gyms [23]. Therefore, considering location as a situation-specific signal could lead to substantial improvements in the recommendation performance. *Time of the day* is another situational signal that could be used for recommendation; for instance, the music a user would like to listen to in mornings differs from those in nights [28]. There are a lot of other situational signals, including but are not limited to, the user's current activity [142], the weather [105], day of the week [60], and the user's mood [97]. It is worth noting that situational features have been proven to be strong signals in improving retrieval performance in search engines [12, 147]. Therefore, we believe that researching and building situation-aware music recommender systems should be one central topic in MRS research.

While several situation-aware MRS already exist, e.g. [11, 23, 67, 74, 116, 142], they commonly exploit only one or very few such situational signals, or are restricted to a certain usage context, e.g., music consumption in a car or in a tourist scenario. Those systems that try to take a more comprehensive view and consider a variety of different signals, on the other hand, suffer from a low number of data instances or users, rendering it very hard to build accurate context models [54]. What is still missing, in our opinion, are (commercial) systems that integrate a variety of situational signals on a very large scale in order to truly understand the listeners needs and intents in any given situation and recommend music accordingly. While we are aware that data availability and privacy

concerns counteract the realization of such systems on a large commercial scale, we believe that MRS will eventually integrate decent multifaceted user models inferred from contextual and situational factors.

3.3 Culture-aware music recommendation

While most humans share an inclination to listen to music, independent on their location or cultural background, the way music is performed, perceived, and interpreted evolves in a culture-specific manner. However, research in MRS seems to be agnostic of this fact. In music information retrieval (MIR) research, on the other hand, cultural aspects have been studied to some extent in recent years, after preceding (and still ongoing) criticisms of the predominance of Western music in this community. Arguably the most comprehensive culture-specific research in this domain has been conducted as part of the CompMusic project,[10] in which five non-Western music traditions have been analyzed in detail in order to advance automatic description of music by emphasizing cultural specificity. The analyzed music traditions included Indian Hindustani and Carnatic [36], Turkish Makam [37], Arab-Andalusian [133], and Beijing Opera [109]. However, the project's focus was on music creation, content analysis, and ethnomusicological aspects rather than on the music consumption side [25, 124, 125]. Recently, analyzing content-based audio features describing rhythm, timbre, harmony, and melody for a corpus of a larger variety of world and folk music with given country information, Panteli et al. found distinct acoustic patterns of the music created in individual countries [103]. They also identified geographical and cultural proximities that are reflected in music features, looking at outliers and misclassifications in a classification experiments using country as target class. For instance, Vietnamese music was often confused with Chinese and Japanese, South African with Botswanese.

In contrast to this — meanwhile quite extensive — work on culture-specific analysis of music traditions, little effort has been made to analyze cultural differences and patterns of music consumption behavior, which is, as we believe, a crucial step to build culture-aware MRS. The few studies investigating such cultural differences include [65], in which Hu and Lee found differences in perception of moods between American and Chinese listeners. By analyzing the music listening behavior of users from 49 countries, Ferwerda et al. found relationships between music listening diversity and Hofstede's cultural dimensions [49, 51]. Skowron et al. used the same dimensions to predict genre preferences of listeners with different cultural backgrounds [130]. Schedl analyzed a large corpus of listening histories created by Last.fm users in 47 countries and identified distinct preference patterns [115]. Further analyses revealed countries closest to what can be considered the global mainstream (e.g., the Netherlands, UK, and Belgium) and countries farthest from it (e.g., China, Iran, and Slovakia). However, all of these works define culture in terms of country borders, which often makes sense, but is sometimes also problematic, for instance in countries with large minorities of inhabitants with different culture.

In our opinion, when building MRS, the analysis of cultural patterns of music consumption behavior, subsequent creation of respective cultural listener models, and their integration into recommender systems are vital steps to improve personalization and serendipity of recommendations. Culture should be defined on various levels though, not only country borders. Other examples include having a joint historical background, speaking the same language, sharing the same beliefs or religion, and differences between urban vs. rural cultures. We believe that MRS which are aware of the cross-cultural differences and similarities in music perception and taste, and are able to recommend music a listener in the same or another culture may like, would substantially benefit both users and providers of MRS.

[10] http://compmusic.upf.edu

4 CONCLUSIONS

In this trends and survey paper, we identified several grand challenges the research field of music recommender systems (MRS) is facing. These are, to the best of our knowledge, in the focus of current research in the area of MRS. We discussed (1) the *cold start problem* of items and users, with its particularities in the music domain, (2) the challenge of *automatic playlist continuation*, which is gaining particular importance due to the recently emerged user request of being recommended musical experiences rather than single tracks [120], and (3) the challenge of holistically *evaluating* music recommender systems, in particular, capturing aspects beyond accuracy.

In addition to the grand challenges, which are currently highly researched, we also presented a visionary outlook of what we believe to be the most interesting future research directions in MRS. In particular, we discussed (1) *psychologically-inspired MRS*, which consider in the recommendation process factors such as listeners' emotion and personality, (2) *situation-aware MRS*, which holistically model contextual and environmental aspects of the music consumption process, infer listener needs and intents, and eventually integrate these models at large scale in the recommendation process, and (3) *culture-aware MRS*, which exploit the fact that music taste highly depends on the cultural background of the listener, where culture can be defined in manifold ways, including historical, political, linguistic, or religious similarities.

We hope that this article helped pinpointing major challenges, highlighting recent trends, and identifying interesting research questions in the area of music recommender systems. Believing that research addressing the discussed challenges and trends will pave the way for the next generation of music recommender systems, we are looking forward to exciting, innovative approaches and systems that improve user satisfaction and experience, rather than just accuracy measures.

5 ACKNOWLEDGEMENTS

We would like to thank all researchers in the fields of recommender systems, information retrieval, music research, and multimedia, with whom we had the pleasure to discuss and collaborate in recent years, and whom in turn influenced and helped shaping this article. Special thanks go to Peter Knees and Fabien Gouyon for the fruitful discussions while preparing the ACM Recommender Systems 2017 tutorial on music recommender systems. We would also like to thank Eelco Wiechert for providing additional pointers to relevant literature. Furthermore, the many personal discussions with

actual users of MRS unveiled important shortcomings of current approaches and in turn were considered in this article.

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2015. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 54.

[2] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alexander Tuzhilin. 2011. Context-aware recommender systems. *AI Magazine* 32 (2011), 67–80. Issue 3.

[3] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17, 6 (June 2005), 734–749. DOI: http://dx.doi.org/10.1109/TKDE.2005.99

[4] Deepak Agarwal and Bee-Chung Chen. 2009. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 19–28.

[5] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender Systems*. Springer, 139–166.

[6] Charu C Aggarwal. 2016. Ensemble-based and hybrid recommender systems. In *Recommender Systems*. Springer, 199–224.

[7] Charu C Aggarwal. 2016. Evaluating Recommender Systems. In *Recommender Systems*. Springer, 225–254.

[8] Masoud Alghoniemy and Ahmed Tewfik. 2001. A Network Flow Model for Playlist Generation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. Tokyo, Japan.

[9] Masoud Alghoniemy and Ahmed H Tewfik. 2000. User-defined music sequence retrieval. In *Proceedings of the eighth ACM international conference on Multimedia*. ACM, 356–358.

[10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval – The Concepts and Technology Behind Search* (2nd ed.). Addison-Wesley, Pearson, Harlow, England.

[11] Linas Baltrunas, Marius Kaminskas, Bernd Ludwig, Omar Moling, Francesco Ricci, Karl-Heinz Lüke, and Roland Schwaiger. 2011. InCarMusic: Context-Aware Music Recommendations in a Car. In *International Conference on Electronic Commerce and Web Technologies (EC-Web)*. Toulouse, France.

[12] Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz. 2011. Inferring and Using Location Metadata to Personalize Web Search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 135–144. DOI: http://dx.doi.org/10.1145/2009916.2009938

[13] Dmitry Bogdanov, Martín Haro, Ferdinand Fuhrmann, Anna Xambó, Emilia Gómez, and Perfecto Herrera. 2013. Semantic Audio Content-based Music Recommendation and Visualization Based on User Preference Examples. *Information Processing & Management* 49, 1 (2013), 13–33.

[14] Dirk Bollen, Bart P. Knijnenburg, Martijn C. Willemsen, and Mark Graus. 2010. Understanding Choice Overload in Recommender Systems. In *Proceedings of the 4th ACM Conference on Recommender Systems*. Barcelona, Spain.

[15] Geoffray Bonnin and Dietmar Jannach. 2015. Automated generation of music playlists: Survey and experiments. *ACM Computing Surveys (CSUR)* 47, 2 (2015), 26.

[16] John S. Breese, David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc, 43–52.

[17] Jerry M. Burger. 2010. *Personality*. Wadsworth Publishing, Belmont, CA., USA.

[18] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[19] Robin Burke. 2007. *Hybrid Web Recommender Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 377–408. DOI: http://dx.doi.org/10.1007/978-3-540-72079-9_12

[20] Iván Cantador and Paolo Cremonesi. 2014. Tutorial on Cross-domain Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14)*. ACM, New York, NY, USA, 401–402. DOI: http://dx.doi.org/10.1145/2645710.2645777

[21] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. 2015. *Cross-Domain Recommender Systems*. Springer US, Boston, MA, 919–959. DOI: http://dx.doi.org/10.1007/978-1-4899-7637-6_27

[22] S. Chen, J.L. Moore, D. Turnbull, and T. Joachims. 2012. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. Beijing, China.

[23] Zhiyong Cheng and Jialie Shen. 2014. Just-for-Me: An Adaptive Personalization System for Location-Aware Social Music Recommendation. In *Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR)*. Glasgow, UK.

[24] Zhiyong Cheng and Jialie Shen. 2016. On effective location-aware music recommendation. *ACM Transactions on Information Systems (TOIS)* 34, 2 (2016), 13.

[25] Olmo Cornelis, Joren Six, Andre Holzapfel, and Marc Leman. 2013. Evaluation and Recommendation of Pulse and Tempo Annotation in Ethnic Music. *Journal of New Music Research* 42, 2 (2013), 131–149. DOI: http://dx.doi.org/10.1080/09298215.2013.812123

[26] Paolo Cremonesi and Massimo Quadrana. 2014. Cross-domain Recommendations Without Overlapping Data: Myth or Reality?. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 297–300. DOI: http://dx.doi.org/10.1145/2645710.2645769

[27] P. Cremonesi, A. Tripodi, and R. Turrin. 2011. Cross-Domain Recommender Systems. In *IEEE 11th International Conference on Data Mining Workshops*. 496–503. DOI: http://dx.doi.org/10.1109/ICDMW.2011.57

[28] Stuart Cunningham, Stephen Caulder, and Vic Grout. 2008. Saturday Night or Fever? Context-Aware Music Playlists. In *Proceedings of the 3rd International Audio Mostly Conference: Sound in Motion*. Piteå, Sweden.

[29] Sally Jo Cunningham, David Bainbridge, and Annette Falconer. 2006. 'More of an Art than a Science': Supporting the Creation of Playlists and Mixes. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*. Victoria, BC, Canada.

[30] Sally Jo Cunningham, David Bainbridge, and Dana Mckay. 2007. Finding New Music: A Diary Study of Everyday Encounters with Novel Songs. In *Proceedings of the 8th International Conference on Music Information Retrieval*. Vienna, Austria, 83–88.

[31] Sally Jo Cunningham, J. Stephen Downie, and David Bainbridge. 2005. "The Pain, The Pain": Modelling Music Information Behavior And The Songs We Hate. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*. London, UK, 474–477.

[32] L. Dey, M. U. Asad, N. Afroz, and R. P. D. Nath. 2014. Emotion extraction from real time chat messenger. In *2014 International Conference on Informatics, Electronics Vision (ICIEV)*. 1–5. DOI: http://dx.doi.org/10.1109/ICIEV.2014.6850785

[33] Justin Donaldson. 2007. A Hybrid Social-acoustic Recommendation System for Popular Music. In *Proceedings of the ACM Conference on Recommender Systems (RecSys)*. Minneapolis, MN, USA, 4.

[34] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. 2011. The yahoo! music dataset and kdd-cup'11. In *Proceedings of the 2011 International Conference on KDD Cup 2011-Volume 18*. JMLR. org, 3–18.

[35] Greg Dunn, Jurgen Wiersema, Jaap Ham, and Lora Aroyo. 2009. Evaluating Interface Variants on Personality Acquisition for Recommender Systems. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: formerly UM and AH (UMAP '09)*. Springer-Verlag, Berlin, Heidelberg, 259–270.

[36] Shrey Dutta and Hema A. Murthy. 2014. Discovering Typical Motifs of a Raga from One-Liners of Songs in Carnatic Music. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipei, Taiwan, 397–402.

[37] Georgi Dzhambazov, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra. 2016. On the Use of Note Onsets for Improved Lyrics-to-audio Alignment in Turkish Makam Music. In *17th International Society for Music Information Retrieval Conference (ISMIR 2016)*. New York, USA.

[38] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent Neural Networks for Emotion Recognition in Video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, New York, NY, USA, 467–474. DOI: http://dx.doi.org/10.1145/2818346.2830596

[39] Hamid Eghbal-zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. 2015. I-Vectors for Timbre-Based Music Similarity and Music Artist Classification. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*. Malaga, Spain.

[40] Mehdi Elahi. 2011. Adaptive active learning in recommender systems. *User Modeling, Adaption and Personalization* (2011), 414–417.

[41] Mehdi Elahi, Matthias Braunhofer, Francesco Ricci, and Marko Tkalcic. 2013. Personality-based active learning for collaborative filtering recommender systems. In *AI* IA 2013: Advances in Artificial Intelligence*. Springer International Publishing, 360–371. DOI: http://dx.doi.org/10.1007/978-3-319-03524-6_31

[42] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2014. Active Learning Strategies for Rating Elicitation in Collaborative Filtering: A System-wide Perspective. *ACM Transactions on Intelligent Systems and Technology* 5, 1, Article 13 (Jan. 2014), 33 pages. DOI: http://dx.doi.org/10.1145/2542182.2542195

[43] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2016. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review* 20 (2016), 29–50.

[44] Asmaa Elbadrawy and George Karypis. 2015. User-specific feature-based similarity models for top-n recommendation of new items. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 33.

[45] Mehmet Erdal, Markus Kächele, and Friedhelm Schwenker. 2016. *Emotion Recognition in Speech with Deep Learning Architectures*. Springer International Publishing, Cham, 298–311. DOI: http://dx.doi.org/10.1007/978-3-319-46182-3_25

[46] Ignacio Fernandez Tobias, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Cantador Ivan. 2016. Alleviating the New User Problem in Collaborative Filtering by Exploiting Personality Information. *User Modeling and User-Adapted Interaction (UMUAI)* 26, Personality in Personalized Systems (2016). DOI: http://dx.doi.org/10.1007/s11257-016-9172-z

[47] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. 2012. Cross-domain recommender systems: A survey of the state of the art. In *Spanish Conference on Information Retrieval*. 24.

[48] Bruce Ferwerda, Mark Graus, Andreu Vall, Marko Tkalčič, and Markus Schedl. 2016. The Influence of Users' Personality Traits on Satisfaction and Attractiveness of Diversified Recommendation Lists. In *Proceedings of the 4th Workshop on Emotions and Personality in Personalized Services (EMPIRE 2016)*. Boston, USA.

[49] Bruce Ferwerda and Markus Schedl. 2016. Investigating the Relationship Between Diversity in Music Consumption Behavior and Cultural Dimensions: A Cross-country Analysis. In *Workshop on Surprise, Opposition, and Obstruction in Adaptive and Personalized Systems*.

[50] Bruce Ferwerda, Markus Schedl, and Marko Tkalčič. 2015. Personality & Emotional States: Understanding Usersfi Music Listening Needs. In *Extended Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP)*. Dublin, Ireland.

[51] Bruce Ferwerda, Andreu Vall, Marko Tkalčič, and Markus Schedl. 2016. Exploring Music Diversity Needs Across Countries. In *Proc. UMAP*.

[52] Bruce Ferwerda, Emily Yang, Markus Schedl, and Marko Tkalčič. 2015. Personality Traits Predict Music Taxonomy Preferences. In *ACM CHI '15 Extended Abstracts on Human Factors in Computing Systems*. Seoul, Republic of Korea.

[53] Arthur Flexer, Dominik Schnitzer, Martin Gasser, and Gerhard Widmer. 2008. Playlist Generation Using Start and End Songs. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*. Philadelphia, PA, USA.

[54] Michael Gillhofer and Markus Schedl. 2015. Iron Maiden While Jogging, Debussy for Dinner? - An Analysis of Music Listening Behavior in Context. In *Proceedings of the 21st International Conference on MultiMedia Modeling (MMM)*. Sydney, Australia.

[55] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[56] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer, Chapter 8, 256–308.

[57] Marc Hassenzahl. 2005. *The Thing and I: Understanding the Relationship Between User and Product*. Springer Netherlands, Dordrecht, 31–42. DOI: http://dx.doi.org/10.1007/1-4020-2967-5_4

[58] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22, 1 (2004), 5–53. DOI: http://dx.doi.org/10.1145/963770.963772

[59] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, 1 (January 2004), 5–53.

[60] P. Herrera, Z. Resa, and M. Sordo. 2010. Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. In *Proceedings of the ACM Conference on Recommender Systems: Workshop on Music Recommendation and Discovery (WOMRAD 2010)*. 7–10.

[61] K. Hevner. 1935. Expression in Music: A Discussion of Experimental Studies and Theories. *Psychological Review* 42 (March 1935). Issue 2.

[62] Rong Hu and Pearl Pu. 2009. A comparative user study on rating vs. personality quiz based preference elicitation methods. In *Proceedings of the 14th international conference on Intelligent user interfaces (IUI '09)*. ACM, New York, NY, USA, 367–372. DOI: http://dx.doi.org/10.1145/1502650.1502702

[63] Rong Hu and Pearl Pu. 2010. A Study on User Perception of Personality-Based Recommender Systems.. In *UMAP* (2010-06-29) *(Lecture Notes in Computer Science)*, Paul De Bra, Alfred Kobsa, and David N. Chin (Eds.), Vol. 6075. Springer, 291–302.

[64] Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems (RecSys '11)*. ACM, New York, NY, USA, 197–204. DOI: http://dx.doi.org/10.1145/2043932.2043969

[65] Xiao Hu and Jin Ha Lee. 2012. A Cross-cultural Study of Music Mood Perception Between American and Chinese Listeners. In *Proc. ISMIR*.

[66] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining*. IEEE, 263–272.

[67] Yajie Hu and Mitsunori Ogihara. 2011. NextOne Player: A Music Recommendation System Based on User Behavior. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. Miami, FL, USA.

[68] A Huq, J.P. Bello, and R. Rowe. 2010. Automated Music Emotion Recognition: A Systematic Evaluation. *Journal of New Music Research* 39, 3 (November 2010), 227–244.

[69] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446. DOI: http://dx.doi.org/10.1145/582415.582418

[70] Oliver John and Sanjay Srivastava. 1999. The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In *Handbook of personality: Theory and research* (second ed.), Lawrence A Pervin and Oliver P John (Eds.). Number 510. Guilford Press, New York, 102–138.

[71] Oliver P. John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In: Handbook of personality: Theory and research. vol. 2, pp. 102 to 138.

[72] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1, Article 2 (Dec. 2016), 42 pages. DOI: http://dx.doi.org/10.1145/2926720

[73] Marius Kaminskas and Francesco Ricci. 2012. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review* 6, 2 (2012), 89–119.

[74] Marius Kaminskas, Francesco Ricci, and Markus Schedl. 2013. Location-aware Music Recommendation Using Auto-Tagging and Hybrid Matching. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*. Hong Kong, China.

[75] John Paul Kelly and Derek Bridge. 2006. Enhancing the diversity of conversational collaborative recommendations: a comparison. *Artificial Intelligence Review* 25, 1 (01 Apr 2006), 79–95. DOI: http://dx.doi.org/10.1007/s10462-007-9023-8

[76] Muhammad Murad Khan, Roliana Ibrahim, and Imran Ghani. 2017. Cross Domain Recommender Systems: A Systematic Literature Review. *ACM Computing Surveys (CSUR)* 50, 3 (2017), 36.

[77] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proc. ISMIR*. 255–266.

[78] Daniel Kluver and Joseph A Konstan. 2014. Evaluating recommender behavior for new users. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 121–128. DOI: http://dx.doi.org/10.1145/2645710.2645742

[79] Peter Knees, Tim Pohle, Markus Schedl, and Gerhard Widmer. 2006. Combining Audio-based Similarity with Web-based Data to Accelerate Automatic Music Playlist Generation. In *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*. Santa Barbara, CA, USA.

[80] P. Knees and M. Schedl. 2016. *Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies*. Springer Berlin Heidelberg. https://books.google.it/books?id=MdRhjwEACAAJ

[81] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[82] Vladimir J Konecni. 1982. Social interaction and musical preference. *The psychology of music* (1982), 497–516.

[83] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (April 2013), 5802fi?!–5805.

[84] M. Kosinski, D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* (March 2013), 2–5. DOI: http://dx.doi.org/10.1073/pnas.1218772110

[85] Fang-Fei Kuo, Meng-Fen Chiang, Man-Kwan Shan, and Suh-Yin Lee. 2005. Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 507–510.

[86] Audrey Laplante. 2014. Improving Music Recommender Systems: What We Can Learn From Research on Music Tastes?. In *15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan.

[87] Audrey Laplante and J. Stephen Downie. 2006. Everyday Life Music Information-Seeking Behaviour of Young Adults. In *Proceedings of the 7th International Conference on Music Information Retrieval*. Victoria (BC), Canada.

[88] Jin Ha Lee. 2011. How Similar is Too Similar? Exploring Users' Perceptions of Similarity in Playlist Evaluation. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. Miami, FL, USA.

[89] Qing Li, Sung Hyon Myaeng, Dong Hai Guan, and Byeong Man Kim. 2005. A probabilistic model for music recommendation considering audio features. In *Asia Information Retrieval Symposium*. Springer, 72–83.

[90] Yanen Li, Jia Hu, ChengXiang Zhai, and Ye Chen. 2010. Improving one-class collaborative filtering by incorporating rich user information. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 959–968.

[91] Nathan N. Liu and Qiang Yang. 2008. EigenRank: a ranking-oriented approach to collaborative filtering. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 83–90. DOI: http://dx.doi.org/10.1145/1390334.1390351

[92] Beth Logan. 2002. Content-based Playlist Generation: Exploratory Experiments. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR)*. Paris, France.

[93] Adam J Lonsdale and Adrian C North. 2011. Why do we listen to music? A uses and gratifications analysis. *British Journal of Psychology* 102, 1 (February 2011), 108–134.

[94] François Maillet, Douglas Eck, Guillaume Desjardins, Paul Lamere, and others. 2009. Steerable Playlist Generation by Learning Song Similarity from Radio Station Playlists.. In *ISMIR*. 345–350.

[95] Brian McFee and Gert Lanckriet. 2012. Hypergraph Models of Playlist Dialects. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*. Porto, Portugal.

[96] Brian McFee and Gert RG Lanckriet. 2011. The Natural Language of Playlists.. In *ISMIR*, Vol. 11. 537–542.

[97] A.C. North and D.J. Hargreaves. 1996. Situational influences on reported musical preference. *Psychomusicology: Music, Mind and Brain* 15, 1-2 (1996), 30–45.

[98] Adrian North and David Hargreaves. 2008. *The social and applied psychology of music*. Oxford University Press.

[99] Adrian C North and David J Hargreaves. 1996. Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition* 15, 1-2 (1996), 30.

[100] Alberto Novello, Martin F. McKinney, and Armin Kohlrausch. 2006. Perceptual Evaluation of Music Similarity. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*. Victoria, BC, Canada.

[101] F Pachet, P Roy, and D Cazaly. 1999. A combinatorial approach to content-based music selection. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, Vol. 1. IEEE, 457–462.

[102] Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *Proceedings of the 8th IEEE International Conference on Data Mining*. IEEE, 502–511.

[103] Maria Panteli, Emmanouil Benetos, and Simon Dixon. 2016. Learning a Feature Space for Similarity in World Music. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*. New York, NY, USA.

[104] Seung-Taek Park and Wei Chu. 2009. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems (RecSys '09)*. ACM, New York, NY, USA, 21–28. DOI: http://dx.doi.org/10.1145/1639714.1639720

[105] T.F. Pettijohn, G.M. Williams, and T.C. Carter. 2010. Music for the Seasons: Seasonal Music Preferences in College Students. *Current Psychology* (2010), 1–18.

[106] Martin Pichl, Eva Zangerle, and Günther Specht. 2015. Towards a context-aware music recommendation approach: What is hidden in the playlist name?. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 1360–1365.

[107] Tim Pohle, Peter Knees, Markus Schedl, Elias Pampalk, and Gerhard Widmer. 2007. "Reinventing the Wheel": A Novel Approach to Music Player Interfaces. *IEEE Transactions on Multimedia* 9 (2007), 567–575. Issue 3.

[108] Peter J. Rentfrow and Samuel D. Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology* 84, 6 (2003), 1236–1256.

[109] Rafael Caro Repetto and Xavier Serra. 2014. Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis. In *15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 313–318.

[110] Gordon Reynolds, Dan Barry, Ted Burke, and Eugene Coyle. 2007. Towards a Personal Automatic Music Playlist Generation Algorithm: The Need for Contextual Information. In *Proceedings of the 2nd International Audio Mostly Conference: Interaction with Sound*. Ilmenau, Germany, 84–89.

[111] Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient Hybridization for Multi-objective Recommender Systems. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. ACM, New York, NY, USA, 19–26. DOI: http://dx.doi.org/10.1145/2365952.2365962

[112] Neil Rubens, Mehdi Elahi, Masashi Sugiyama, and Dain Kaplan. 2015. Active Learning in Recommender Systems. In *Recommender Systems Handbook - chapter 24: Recommending Active Learning*. Springer US, 809–846.

[113] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[114] Thomas Schäfer and Claudia Mehlhorn. 2017. Can Personality Traits Predict Musical Style Preferences? A Meta-Analysis. *Personality and Individual Differences* 116 (2017), 265–273. DOI: http://dx.doi.org/https://doi.org/10.1016/j.paid.2017.04.061

[115] Markus Schedl. 2017. Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. *International Journal of Multimedia Information Retrieval* 6, 1 (2017), 71–84. DOI: http://dx.doi.org/10.1007/s13735-017-0118-y

[116] Markus Schedl, Georg Breitschopf, and Bogdan Ionescu. 2014. Mobile Music Genius: Reggae at the Beach, Metal on a Friday Night?. In *Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR)*. Glasgow, UK.

[117] Markus Schedl, Arthur Flexer, and Julián Urbano. 2013. The Neglected User in Music Information Retrieval Research. *Journal of Intelligent Information Systems* (July 2013).

[118] Markus Schedl, Emilia Gómez, Erika S. Trent, Marko Tkalčič, Hamid Eghbal-Zadeh, and Agustin Martorell. 2017. On the Interrelation between Listener Characteristics and the Perception of Emotions in Classical Orchestra Music. *IEEE Transactions on Affective Computing* (2017).

[119] Markus Schedl, David Hauger, and Dominik Schnitzer. 2012. A Model for Serendipitous Music Retrieval. In *Proceedings of the 2nd Workshop on Context-awareness in Retrieval and Recommendation (CaRR)*. Lisbon, Portugal.

[120] Markus Schedl, Peter Knees, and Fabien Gouyon. 2017. New Paths in Music Recommender Systems Research. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017)*. Como, Italy.

[121] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. 2015. Music Recommender Systems. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer, Chapter 13, 453–492.

[122] Markus Schedl, Mark Melenhorst, Cynthia C.S. Liem, Agustín Martorell, 'Oscar Mayor, and Marko Tkalčič. 2016. A Personality-based Adaptive System for Visualizing Classical Music Performances. In *Proceedings of the 7th ACM Multimedia Systems Conference (MMSys)*. Klagenfurt, Austria.

[123] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 253–260. DOI: http://dx.doi.org/10.1145/564376.564421

[124] Xavier Serra. 2014. Computational Approaches to the Art Music Traditions of India and Turkey.

[125] Xavier Serra. 2014. Creating Research Corpora for the Computational Study of Music: the case of the CompMusic Project. In *AES 53rd International Conference on Semantic Audio*. AES, AES, London, UK, 1–9.

[126] Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter Knees. 2010. Using Block-Level Features for Genre Classification, Tag Classification and Music Similarity Estimation. In *Extended Abstract to the Music Information Retrieval Evaluation eXchange (MIREX 2010) / 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*. Utrecht, the Netherlands.

[127] Klaus Seyerlehner, Gerhard Widmer, Markus Schedl, and Peter Knees. 2010. Automatic Music Tag Classification based on Block-Level Features. In *Proceedings of the 7th Sound and Music Computing Conference (SMC)*. Barcelona, Spain.

[128] Bo Shao, Dingding Wang, Tao Li, and Mitsunori Ogihara. 2009. Music Recommendation Based on Acoustic Features and User Access Patterns. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 8 (2009), 1602–1611.

[129] Marcin Skowron, Bruce Ferwerda, Marko Tkalčič, and Markus Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *Proceedings of the 25th International World Wide Web Conference (WWW)*. Montreal, Canada.

[130] M. Skowron, F. Lemmerich, Bruce Ferwerda, and Markus Schedl. 2017. Predicting Genre Preferences from Cultural and Socio-economic Factors for Music Retrieval. In *Proc. ECIR*.

[131] Malcolm Slaney and William White. 2006. Measuring playlist diversity for recommendation systems. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, 77–82.

[132] Barry Smyth and Paul McClave. 2001. Similarity vs. Diversity. In *Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development (ICCBR '01)*. SpringerVerlag, London, UK, 347–361. http://dl.acm.org/citation.cfm?id=646268.758890

[133] Mohamed Sordo, Amin Chaachoo, and Xavier Serra. 2014. Creating Corpora for Computational Research in Arab-Andalusian Music. In *1st International Workshop on Digital Libraries for Musicology*. London, UK, 1–3. DOI: http://dx.doi.org/10.1145/2660168.2660182

[134] Kirsten Swearingen and Rashmi Sinha. 2001. Beyond algorithms: An HCI perspective on recommender systems. In *ACM SIGIR 2001 Workshop on Recommender Systems*, Vol. 13. 1–11.

[135] Nava Tintarev, Christoph Lofi, and Cynthia C.S. Liem. 2017. Sequences of Diverse Song Recommendations: An Exploratory Study in a Commercial System. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 391–392. DOI: http://dx.doi.org/10.1145/3079628.3079633

[136] Marko Tkalcic, Andrej Kosir, and Jurij Tasic. 2013. The LDOS-PerAff-1 corpus of facial-expression video clips with affective, personality and user-interaction metadata. *Journal on Multimodal User Interfaces* 7, 1-2 (2013), 143–155. DOI: http://dx.doi.org/10.1007/s12193-012-0107-7

[137] Marko Tkalčič, Daniele Quercia, and Sabine Graf. 2016. Preface to the special issue on personality in personalized systems. *User Modeling and User-Adapted Interaction* 26, 2 (June 2016), 103–107. DOI: http://dx.doi.org/10.1007/s11257-016-9175-9

[138] Alexandra Uitdenbogerd and R Schyndel. 2002. A review of factors affecting music recommender success. In *ISMIR 2002, 3rd International Conference on Music Information Retrieval*. IRCAM-Centre Pompidou, 204–208.

[139] Andreu Vall, Massimo Quadrana, Markus Schedl, Gerhard Widmer, and Paolo Cremonesi. 2017. The Importance of Song Context in Music Playlists. In *Proceedings of the Poster Track of the 11th ACM Conference on Recommender Systems (RecSys)*. Como, Italy.

[140] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, Redundancy and Size-awareness in Genre Diversity for Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*. ACM, New York, NY, USA, 209–216. DOI: http://dx.doi.org/10.1145/2645710.2645743

[141] Saúl Vargas and Pablo Castells. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys)*. Chicago, IL, USA, 8.

[142] Xinxi Wang, David Rosenblum, and Ye Wang. 2012. Context-aware Mobile Music Recommendation for Daily Activities. In *Proceedings of the 20th ACM International Conference on Multimedia*. ACM, Nara, Japan, 99–108.

[143] Markus Weimer, Alexandros Karatzoglou, and Alex Smola. 2008. Adaptive collaborative filtering. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*. ACM, New York, NY, USA, 275–282. DOI: http://dx.doi.org/10.1145/1454008.1454050

[144] Yi-Hsuan Yang and Homer H. Chen. 2011. *Music Emotion Recognition*. CRC Press.

[145] Yi-Hsuan Yang and Homer H. Chen. 2013. Machine Recognition of Music Emotion: A Review. *Transactions on Intelligent Systems and Technology* 3, 3 (May 2013).

[146] Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. 2006. Hybrid Collaborative and Content-based Music Recommendation Using Probabilistic Model with Latent User Preferences.. In *ISMIR*, Vol. 6. 7th.

[147] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational Context for Ranking in Personal Search. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1531–1540. DOI: http://dx.doi.org/10.1145/3038912.3052648

[148] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* 8, 4 (2008), 494.

[149] Zhang, Yuan Cao and O Seaghdha, Diarmuid and Quercia, Daniele and Jambor, Tamas. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*. Seattle, WA, USA.

[150] Elena Zheleva, John Guiver, Eduarda Mendes Rodrigues, and Nataša Milić-Frayling. 2010. Statistical Models of Music-listening Sessions in Social Media. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*. Raleigh, NC, USA, 1019–1028.

[151] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.

[152] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on the World Wide Web*. ACM, 22–32.