Iron Maiden while jogging, Debussy for dinner?

An analysis of music listening behavior in context

Michael Gillhofer and Markus Schedl

Johannes Kepler University
Linz, Austria
http://www.cp.jku.at

Abstract. Contextual information of the listener is only slowly being integrated into music retrieval and recommendation systems. Given the enormous rise in mobile music consumption and the many sensors integrated into today's smart-phones, at the same time, an unprecedented source for user context data of different kinds is becoming available.
Equipped with a smart-phone application, which had been developed to monitor contextual aspects of users when listening to music, we collected contextual data of listening events for 48 users. About 100 different user features, in addition to music meta-data have been recorded.
In this paper, we analyze the relationship between aspects of the *user context* and *music listening preference*. The goals are to assess (i) whether user context factors allow predicting the song, artist, mood, or genre of a listened track, and (ii) which contextual aspects are most promising for an accurate prediction. To this end, we investigate various classifiers to learn relations between user context aspects and music meta-data. We show that the user context allows to predict artist and genre to some extent, but can hardly be used for song or mood prediction. Our study further reveals that the level of listening activity has little influence on the accuracy of predictions.

1 Introduction

Ever increasing amounts of music available on mobile devices, such as smartphones, demand for intelligent ways to access music collections. In particular mobile music consumption, for instance, via audio streaming services, has been spiraling during the past couple of years. However, accessing songs in mobile music collections is still performed either via simple meta-data filtering and search or via standard collaborative filtering, both ignoring important characteristics of the users, such as their current activity or location. Searching by meta-data performs well when the user has a specific information or entertainment need in mind, collaborative filtering when the user wants to listen to music judged similar by like-minded users. However, these methods do not encourage serendipitous experiences when discovering a music collection.

Integrating the user context in approaches to music retrieval and recommendation has been proposed as a possible solution to remedy the aforementioned shortcomings [15, 19]. Building user-aware music access systems, however, first requires to investigate which characteristics of the listeners (both intrinsic and external) influence their music taste. This paper hence studies a wide variety of user context attributes and assesses how well they perform to predict music taste at various levels: artist, track, genre, and mood. The dataset used in this study has been gathered via a mobile music player that offers automated adaptation of playlists, dependent on the user context [9].

In the remainder, related work is reviewed (Section 2) and the data acquisition process is detailed (Section 3). Subsequently, the experimental setup is defined and classification results are presented, for individual users, for groups of users, and using different categories of features (Section 4). To round off, conclusions are drawn and future work is pointed out (Section 5).

2 Related Work

Context-aware approaches to music retrieval and applications for music access, which take into account the user in a *comprehensive* way, have not been seen before the past few years, to the best of our knowledge. Related work on context-aware music retrieval and recommendation hence differs considerably in how the user context is defined, gathered, and incorporated [19]. Some approaches rely solely on one or a few aspects, such as temporal features [3], listening history and weather conditions [14], while others model the user context in a more comprehensive manner.

The first available **user-aware music access systems** monitored just a particular type of user characteristics to address a specific music consumption scenario. A frequently targeted scenario was to adapt the music to the pace of a jogger, using his pulse rate [2, 17, 16]. However, almost all proposed systems required additional hardware for context logging [6–8].

A few recent approaches model the user via a larger variety of factors, but address only a particular listening scenario. For instance, Kaminskas and Ricci [12] propose a system that matches tags describing a particular place or point of interest with tags describing music. Employing text-based similarity measures between the lists of tags, they target location-based music recommendation. The approach is later extended in [13], where tags for unknown music are automatically learned via a music auto-tagger, from input of a user questionnaire. Baltrunas et al. [1] propose an approach to context-aware music recommendation while driving. The authors take into account eight different contextual factors, such as driving style, mood, road type, weather, and traffic conditions, which they gather via a questionnaire and use to extend a matrix factorization model. In contrast to these works, the mobile music player through which the data analyzed here has been collected logs the listening context in a comprehensive and unobtrusive manner.

Other recently proposed systems for user-aware music recommendation include “NextOne” and “Just-for-me”, the former proposed by Hu and Ogihara [11], the latter by Cheng and Shen [5]. The NextOne player models the music recommendation problem under five perspectives: music genre, release year, user’s favorite music, “freshness” referring to old songs that a user almost forgot and that should be recovered, and temporal aspects per day and week. These five factors are then individually weighted and aggregated to obtain the final recommendations. In the Just-for-me system, the user's location is monitored, music content analysis is performed to obtain audio features, and global music popularity trends are inferred from microblogs. The authors then extend a topic modeling approach to integrate the diverse aspects and in turn offer music recommendations based on audio content, location, listening history, and overall popularity.

For what concerns **user studies on the relation of user-specific aspects and music taste**, the body of scientific work is quite sparse. Cunningham et al. [6] present a study that investigates if and how various factors relate to music taste (e.g., human movement, emotional status, and external factors such as temperature and lightning conditions). Based on the findings, the authors employ a fuzzy logic model to create playlists. Although related to the study at hand, Cunningham et al.'s work has several limitations, foremost (i) the artificial setting because a stationary controller is used to record human movement and (ii) the limitation to eight songs. The study at hand, in contrast, employs a far more flexible setup that monitors music preference and user context in the real world and in an unobtrusive way.

Another study related to the work at hand was performed by Yang and Liu [21], who investigate the interrelation of user mood and music emotion. To this end, Yang and Liu identify user moods from blogs posted on LiveJournal[1] and relate them to music mentioned in the same posting. They show that user mood can be predicted more accurately from the user context, assumed to be reflected in the textual content of the postings, than from audio features extracted from the music mentioned in the postings. While their study focuses on predicting mood from music listening events, our goal is to predict music taste from a wide range of user characteristics, including mood.

3 Data Acquisition

A recently developed smart-phone application called "Mobile Music Genius" [18] allows to monitor the context of the user while listening to music. We analyze the dataset which has been recorded by this application from January to July 2013, foremost for students from the Johannes Kepler University Linz, Austria. It consists of 7628 individual samples from 48 unique persons. We managed to identify 4149 different tracks from 1169 unique artists. As genre and mood data has not been directly recorded by the application, we queried the Last.fm API[2] to obtain this additional information. Unfortunately, the Last.fm data turned out to be quite noisy or not available at all. We were nevertheless able to identify 24 different genres and 70 different moods by matching the Last.fm tags to a dictionary of genres and moods gathered from Freebase[3]. This matching resulted

[1] `http://www.livejournal.com/`

[2] `http://www.lastfm.at/api/`

[3] `http://www.freebase.com/`

Category	**Attributes**
Time	day of week (N), hour of day (N)
Location	provider (C), latitude (C), longitude (C), accuracy (N), altitude (N)
Weather	temperature (N), wind direction (N), wind speed (N), precipitation (N), humidity (N), visibility (N), pressure (N), cloud cover (N), weather code (N)
Device	battery level (N), battery status (N), available internal/external storage (N), volume settings (N), audio output mode (C)
Phone	service state (C), roaming (C), signal strength (N), GSM indicator (N), network type (N)
Task	up to ten recently used tasks/apps (C), screen on/off (C), docking mode (C)

Network	*mobile network*: available (C), connected (C);
	active network: type (C), subtype (C), roaming (C);
	Bluetooth: available (C), enabled (C);
	Wi-Fi: enabled (C), available (C), connected (C), BSSID (C), SSID (C), IP (N), link speed (N), RSSI (N)

Ambient	mean and standard deviation of all attributes: light (N), proximity (N), temperature (N), pressure (N), noise (N)
Motion	mean and standard deviation of acceleration force (N) and rate of rotation (C); orientation of user (N), orientation of device (C)
Player	repeat mode (C), shuffle mode (C), automated playlist modification mode (C), *sound effects*: equalizer present (C), equalizer enabled (C), bass boost enabled (C), bass boost strength (N), virtualizer enabled (C), virtualizer strength (N), reverb enabled (C), reverb strength (N)
Activity	activity (C), mood (N)

Table 1. Monitored user attributes and their type (N=numerical, C=categorical).

in 4246 and 2731 samples, respectively, for genre and mood. The most frequent genres in the dataset are rock (1183 instances), electronic (392), folk (274), metal (224), and hiphop (184). The most frequent moods are party (319), epic (312), sexy (218), happy (154), and sad (153). Arguably, not all of the Freebase mood tags would be considered as mood in a psychological interpretation, but we did not want to artificially restrict the mood data from Freebase and Last.fm. In cases where an artist or song was assigned several genre or mood labels, we selected the one with highest weight according to Last.fm, since we consider a single-label classification problem.

Table 2 summarizes the basic statistics of our dataset for different meta-data levels: the number of instances or data points, the number of unique classes, and the number of users for whom data was available. Table 3 additionally shows per-user-statistics. Notably, the average number of genres per user is quite high (5.14). This means that participants in the study showed a diverse music taste. Figure 1 shows the different activity levels of users. We see a few users have recorded lots of samples. However, compared to them, the majority have been fairly inactive.

	Instances	Classes	Users
Artists	7628	1169	48
Genres	4246	24	45
Moods	2731	70	45
Tracks	7628	4149	48

Table 2. Basic properties of the recorded dataset: number of different data instances, number of unique classes, and number of unique users.

Property	Mean	Med.	SD	Min.	Max.
Artists per user	27.88	13	33.68	1	158
Genres per user	5.14	4	3.84	1	16
Moods per user	9.91	9	9.03	1	36
Titles per user	89.16	46	96.66	1	387

Table 3. Arithmetic mean, median, standard deviation, minimum and maximum, per user and class.

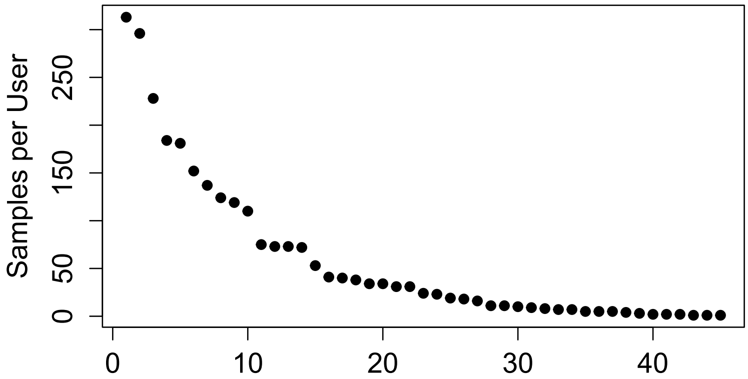

Fig. 1. Distribution of number of data instances per user, in descending order.

4 Predicting the User's Music Taste

Addressing the first research question of whether user context factors allow to predict song, artist, genre, or mood, we performed classification experiments, using standard machine learning algorithms from the *Weka* [10] environment. These were *IBk* (a k-nearest neighbor, instance-based classifier), *J48* (a decision tree learner), *JRip* (a rule learner), *Random Forests*, and *ZeroR*. The last one just predicts the most frequent class among the given training samples, and is therefore used as a baseline. Optimizing the classifiers' parameters has been investigated, but we could not make out a single setting which yielded a substantially better classification accuracy across multiple experiments, hence we used the default configurations in the experiments reported in the following. By

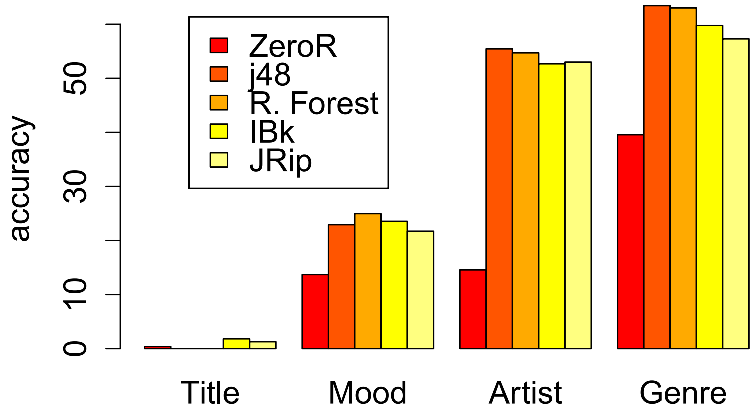

Fig. 2. Accuracy (in %) of classifications using all features.

performing 10-fold cross validation, we estimated the average accuracy of the classifiers' predictions.

The results evidence differences between classifiers. But no single classifier was able to outperform all others in multiple tasks (cf. Figure 2, which shows accuracies for the different classifiers in %). We also could not make out a classifier besides *ZeroR* which yields worse results than the others. Except for that, results vary only up to 10% in accuracy, depending on the experiment.

The average performance of the four non-baseline classifiers vary strongly, however, for different classification tasks: predicting genre, mood, artist, and track. Although our dataset consists of 1169 unique classes for the *artist* classification task, the classifiers managed to correctly predict about 55% of the samples, a remarkable result considering the many classes and 13% accuracy when using majority voting. The *genre* prediction results are quite good as well, since all classifiers obtained a decent accuracy of about 61% correctly predicted samples. Even given the 39% accuracy achieved by the *ZeroR* baseline, this result is remarkable. Predicting the *mood* of music succeeded on average for only about 23% of the samples. It seems that information required to accurately relate user context to music mood labels is not included in the recorded aspects. The last classification task was *title* prediction, which did not work at all. Only about 1.5% of samples have been assigned the correct title. This is not a surprise as the average playcount per title is only 1.83, thus rendering the training of classifiers almost impossible for a large number of users.

To investigate whether prediction accuracy varies for different groups of users and categories of features, we created subsets of the data in different ways:

1. for each user individually,
2. for groups of users according to their activity, and
3. for categories of features.

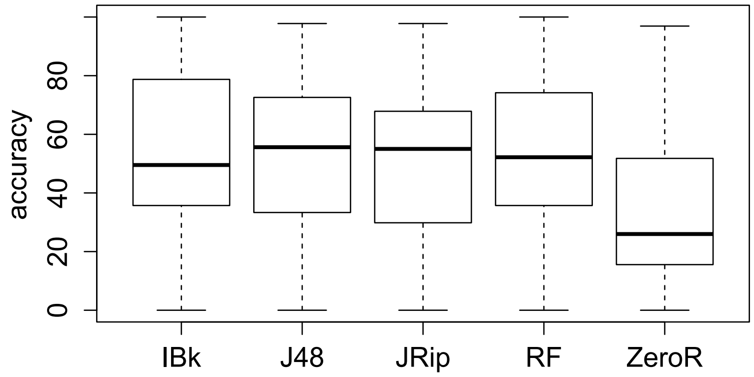

Fig. 3. Boxplot showing accuracy (in %) for each user-specific dataset on the artist prediction task.

4.1 Individual users

We prepared datasets in a way which required each included user to have listened to a minimum of four different tracks. Seven users did not meet this requirement and have been sorted out. We then ran experiments, using as training set only the individual user's data. Experiments were conducted again using 10-fold cross validation. For users for whom the number of samples were below 10, we performed leave-one-out cross-validation.

Figure 3 shows the distribution of the classification results for individual users in the *artist* prediction task, for each used classifier. In this boxplot, the central thick line marks the median, the upper and lower edges of the box mark the 0.25 and 0.75 percentiles, respectively, and the whiskers extend to the highest and lowest values which are not consideres outliers. We see that on average classification works considerably well, but the accuracy varies substantially between different users. We found this behavior for all four classification tasks, but investigate only the *artist* prediction task further, because results were most significant here.

By investigating the type of users for which the number of correct predictions is low, we found that they seem to have a fairly static context while listening to music. The users showing better predictability tend to listen to music in many different contexts. Recommendation systems should thus distinguish between these groups. Separating these two groups may be performed by computing the entropy of users' context features.

4.2 User groups with respect to listening activity

Assuming that not only the diversity of the user context influences the quality of prediction results, as indicated above, but also the number of listening events recorded play an important role, we compared different types of users. To this end, we first sorted the users according to their number of listening events, in

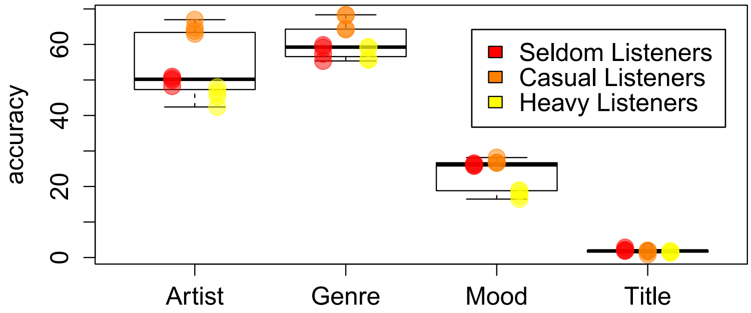

Fig. 4. Accuracies (in %) of all three user groups and all four non-baseline classifiers, for the four classification tasks. Boxplots show the aggregates of the results over all user groups, for each classifier.

descending order. We then divided the dataset into three groups of users: *heavy listeners*, *casual listeners*, and *seldom listeners*. Each group was constructed to cover about one third of all available samples. Hence, the *heavy* group only contains 4 different users, the *casual* group 8, and the *seldom* group the remaining 36 users. The choice of using three groups and accumulated numbers of data instances to separate them was motivated by earlier work on assessing differences in activity or popularity, respectively, between users or artists. To this end, artists or users are typically categorized into three disjoint groups [4, 20].

The classification results for each task are illustrated in Figure 4. We see relatively narrow boxplots for *genre*, *mood*, and *title* predictions, contrasting the results of the *artist* task. We looked deeper into the data and found a cluster of a single artist which corresponds to 18% of all samples within the casual listener group. Therefore, classification of this group seems easier, which results in a higher average accuracy of about 65% with non-baseline classifiers. A similar pattern was found in the genre prediction task, again for the casual listener group. Here, a single genre corresponds to 41% of all samples, which simplifies classification, although the impact is less pronounced. The remaining variability in each classification task can partly be explained by differences of the used classifiers. We conclude that the user's listening activity has only a small influence on the classification results, as long as the user context data is diverse enough.

4.3 Feature categories

Table 1 displays all user aspects under consideration. Each feature was categorized already in [18] into one of the following 11 groups: *Time*, *Location*, *Weather*, *Device*, *Phone*, *Task*, *Network*, *Ambient*, *Motion*, *Player*, and *Activity*. For example, the features *day of week* and *hour of day* both belong to category *Time*. By using only one category for predicting the music listening behavior in our classification tasks, it becomes possible to estimate the importance of the respective kinds of features.

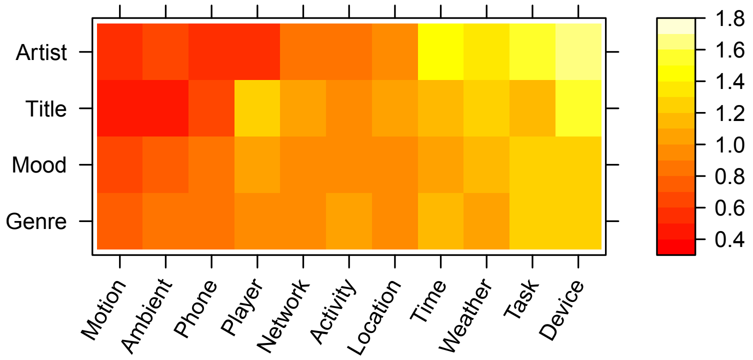

Fig. 5. The relative importance of each feature group *compared to the mean classification result* (achieved over all individual feature categories), per classification task.

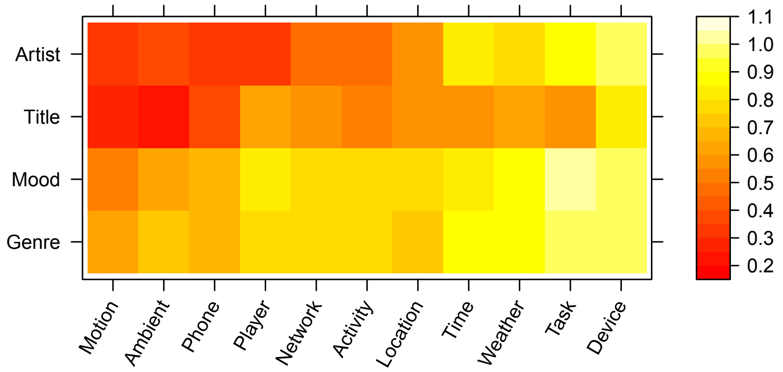

Fig. 6. The relative importance of each feature group *compared to the results obtained including all features*, per classification task.

We trained all classifiers for each feature group and classification task. The results are shown in Figures 5 and 6. We ordered the feature categories from left to right in increasing order according to their value for classification. Each colored box in the matrix represents the average relative performance of the respective category and class, among all four used non-baseline classifiers. Performance is measured in terms of accuracy. In Figure 5, performance values for a particular combination of feature group and classification task (one box) are relative to the mean of the achieved accuracy over all feature groups for that classification task (mean of the respective row of boxes). Performance values reported in Figure 6 for a particular feature group and classification task represent the relative accuracy of that combination, when compared to accuracy obtained by a classifier that exploits all available features.

Therefore, a neutral shade of orange in Figure 5 represents an average importance, whereas darker shades of red indicate a less important group. Consequently, the brighter the shade, the more useful information is contained within this feature group. We see that there are significant differences in the importance of groups. Interestingly, the *Player* feature category can be considered an outlier when it comes to song prediction. Although this feature category might be presumed to be a rather weak indicator, it seems to hold quite valuable information about the title. This could mean that listeners adjust player settings, such as the repeat mode, on certain songs more frequently than on others.

Figure 6 on the other hand shows the relative importance of feature groups compared to the classification accuracy using all features. Hence, a red box indicates an accuracy of only 20-30% of the accuracy achievable using all features, while a bright yellow shade indicates high performance. Therefore, we observe that *Device*, *Task*, *Weather*, and *Time* features contain almost the same amount of information as all features combined. By adding more features, we are not able to increase classification accuracy. Being in line with other research on context-aware systems, the good performance of temporal and weather features is expected. However, also the other tasks running on the user's device while using the music player seem to play a crucial role. In particular, users may prefer certain genres and artists when running a fitness app, but others when checking mails or writing instant messages. Quite surprisingly, device-related aspects are overall most important. A possible explanation is that they typically change very slowly, thus capture the general music taste of the user better than any other aspect.

5 Conclusion and Future Work

We presented a detailed analysis of user context features for the task of predicting music listening behavior, investigating the classes track, artist, genre, and mood. We found substantial differences in classification accuracy, depending on the class. *Genre* classification yielded a remarkable 60% accuracy. *Artist* classification achieved 55% accuracy. Significantly worse results were obtained in the *mood* classification task (25% accuracy) and in particular for the *track* class (1.5% accuracy). Analyzing different groups of users, we found that accuracy is not stable across users, in particular, varies with respect to diversity in user context features. Furthermore, no strong evidence for a correlation between listening activity (number of listening events of a user) and prediction accuracy, for any of the classification tasks, could be made out. We also managed to identify an importance ranking of user context features. Features related to applications running on the device, weather, time, and location turned out to be of particular importance to predict music preference. We further plan to investigate more sophisticated feature selection techniques.

Based on these results, we will elaborate context-aware music recommendation approaches that incorporate the findings presented here. In particular, this study evidences that the diversity of situations or contexts in which a user consumes music has a high impact on the performance of the predictions, and likely in turn also on the performance of corresponding music recommenders. Approaches that incorporate this knowledge along with information about the importance of particular context features should thus be capable to improve over existing solutions.

A possible limitation of the study at hand is the user data it is based upon. In particular, we cannot guarantee that the recruited participants from which we recorded data do correspond to the average music listener, as we required them to have an *Android* device and listen to local music. The user set is also heavily biased towards Austrian students. Although we believe that results are representative, a larger dataset of more and more diverse participants should be created to base future experiments on.

6 Acknowledgments

This research is supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the project "Performances as Highly Enriched aNd Interactive Concert eXperiences" (PHENICX), no. 601166, and by the Austrian Science Fund (FWF): P22856 and P25655.

References

1. L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, K.-H. Lüke, and R. Schwaiger. InCarMusic: Context-Aware Music Recommendations in a Car. In *Proceedings of the International Conference on Electronic Commerce and Web Technologies (EC-Web)*, Toulouse, France, 2011.
2. J. T. Biehl, P. D. Adamczyk, and B. P. Bailey. DJogger: A Mobile Dynamic Music Device. In *CHI 2006: Extended Abstracts on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006.
3. T. Cebrián, M. Planagumà, P. Villegas, and X. Amatriain. Music Recommendations with Temporal Context Awareness. In *Proceedings of the 4th ACM Conference on Recommender Systems*, Barcelona, Spain, 2010.
4. O. Celma. *Music Recommendation and Discovery – The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, Berlin, Heidelberg, Germany, 2010.
5. Z. Cheng and J. Shen. Just-for-Me: An Adaptive Personalization System for Location-Aware Social Music Recommendation. In *Proceedings of the 2014 ACM International Conference on Multimedia Retrieval (ICMR)*, Glasgow, UK, April 2014.
6. S. Cunningham, S. Caulder, and V. Grout. Saturday Night or Fever? Context-Aware Music Playlists. In *Proceedings of the 3rd International Audio Mostly Conference of Sound in Motion*, Piteå, Sweden, October 2008.
7. S. Dornbush, J. English, T. Oates, Z. Segall, and A. Joshi. XPod: A Human Activity Aware Learning Mobile Music Player. In *Proceedings of the IJCAI 2007 Workshop on Ambient Intelligence*, 2007.
8. G. T. Elliott and B. Tomlinson. Personalsoundtrack: Context-aware playlists that adapt to user pace. In *CHI 2006: Extended Abstracts on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006.
9. Georg Breitschopf. Personalized, context-aware music playlist generation on mobile devices. Master's thesis, JKU, Aug. 2013.

10. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18, Nov. 2009.
11. Y. Hu and M. Ogihara. NextOne Player: A Music Recommendation System Based on User Behavior. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, USA, October 2011.
12. M. Kaminskas and F. Ricci. Location-Adapted Music Recommendation Using Tags. In J. Konstan, R. Conejo, J. Marzo, and N. Oliver, editors, *User Modeling, Adaption and Personalization*, volume 6787 of *Lecture Notes in Computer Science*, pages 183–194. Springer Berlin / Heidelberg, 2011.
13. M. Kaminskas, F. Ricci, and M. Schedl. Location-aware Music Recommendation Using Auto-Tagging and Hybrid Matching. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*, Hong Kong, China, October 2013.
14. J. S. Lee and J. C. Lee. Context Awareness by Case-Based Reasoning in a Music Recommendation System. In H. Ichikawa, W.-D. Cho, I. Satoh, and H. Youn, editors, *Ubiquitous Computing Systems*, volume 4836 of *Lecture Notes in Computer Science*, pages 45–58. Springer Berlin / Heidelberg, 2007.
15. C. C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic. The Need for Music Information Retrieval with User-centered and Multimodal Strategies. In *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, Scottsdale, AZ, USA, November 2011.
16. H. Liu and J. H. M. Rauterberg. Music Playlist Recommendation Based on User Heartbeat and Music Preference. In *Proc. 4th Int'l Conf. on Computer Technology and Development (ICCTD)*, Bangkok, Thailand, 2009.
17. B. Moens, L. van Noorden, and M. Leman. D-Jogger: Syncing Music with Walking. In *Proceedings of the 7th Sound and Music Computing Conf. (SMC)*, Barcelona, Spain, 2010.
18. M. Schedl, G. Breitschopf, and B. Ionescu. Mobile Music Genius: Reggae at the Beach, Metal on a Friday Night? In *Proceedings of the 2014 ACM International Conference on Multimedia Retrieval (ICMR)*, Glasgow, UK, April 02-04 2014.
19. M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41:523–539, December 2013.
20. M. Schedl, D. Hauger, and J. Urbano. Harvesting microblogs for contextual music similarity estimation — a co-occurrence-based framework. *Multimedia Systems*, May 2013.
21. Y.-H. Yang and J.-Y. Liu. Quantitative Study of Music Listening Behavior in a Social and Affective Context. *IEEE Transactions on Multimedia*, 15(6):1304–1315, October 2013.