Understanding Playlist Creation on Music Streaming Platforms

Martin Pichl, Eva Zangerle and Günther Specht
Databases and Information Systems
Department of Computer Science at the University of Innsbruck, Austria
Email: firstname.lastname@uibk.ac.at

***Abstract*—Music streaming platforms enable people to access millions of tracks using computers and mobile devices. The latter allow users consume different music during different activities. Both, the sheer amount of music and the mobile access to music makes music organization an interesting topic for multimedia researchers. Assisting users to organize their music and make the music they like easily available in the right moment, contributes to increased usability of music streaming platforms. To get a deeper understanding of how users organize music nowadays, we analyze user-created playlists crawled from the music streaming platform Spotify. Using this new data set we find an explanation of differences in the playlists using audio features and based on this compute playlist clusters. We find that 91% of all users create at least one playlist in the "feel good music"-cluster and classical music or rap music can be considered as niche music with respect to the number of playlists, however not as niche music when considering the number of users. To foster research in this field, we make our analysis tool publicly available.**

***Keywords*-Music Information Retrieval; Data Acquisition; Data Analysis; User-generated Content**

I. Introduction

In the last decade, new technologies have paved way for new distribution channels for digital content, e.g., music streaming platforms like Spotify[1] or Apple Music[2]. At the same time, mobile devices as smartphones or tablets enable their users to access millions of tracks on those streaming platforms in various situations throughout the whole day. These developments make music organization and along with that, context-aware music recommendation, a highly interesting topic: the challenge for the users is to find music they like in the overwhelming variety of music offered by music streaming platforms. In principle, users need to navigate through their music collection to find the music they aim to listen to during different activities or situations [1]. In order to assist users in browsing these possibly extensive collections, streaming platforms heavily rely on recommender systems, but also on human editors. A deeper understanding for the characteristics of playlists and how users create and maintain their playlists can naturally contribute to more personalized and better recommendations.

In the field of music listening behavior analyses and recommender systems, social media platforms have been exploited to gather relevant data for such analyses. Nowadays, a substantial amount of people share what they are listening to at the moment using so-called #nowplayling tweets on Twitter. This makes Twitter, which is the world's leading micro-blogging platform serving 320 million active users[3], a valuable data source. Twitter has already been exploited for various analyses of user listening behavior [2], [3] as well as for recommender systems [4]–[6]. Earlier, automatic playlist generation, as a form of music recommendation, was studied intensively [7]–[12]. Slaney and White found that people prefer different types of music and thus also create playlists biased to this type of music [13]. Furthermore, Cunningham et al. have shown that people categorize music after the intended use [14]. Complementary to this, Kamalzadeh et al. found that people categorize music by activities and/or the mood in their music libraries [1].

In contrast to the well-researched field of automatic playlist generation, we aim to deepen our understanding for the characteristics of playlists created by human users and hence, shift our focus from automatic playlist generation to the analysis of playlists. To conduct this study, we require a data set containing information about users and their playlists. In a previous analysis we found that a substantial portion of so-called #nowplaying tweets refer to Spotify [15]. Along these lines, we create a data set containing Spotify users and their playlists. In total, we base our analyses on 1,133 users and their 18,146 playlists. We are particularly interested in studying the musical attributes of the tracks forming up different playlists, therefore we enrich this data set with music content data crawled from the Echo Nest platform[4]. Our analyses based on this data set are particularly driven by the following research questions (RQ):

- RQ1: How can we observe and explain acoustical differences between playlists using clustering techniques?
- RQ2: How do users utilize playlists of different types to organize their music?

The main contribution of this work is that it is the first to analyze the playlist generation behavior of Spotify users. Further, we provide the first data set containing playlist information gathered from Spotify. We find that using a Principal Component Analysis, we are able to explain differences using content-based music features. When clustering playlists into

[1] http://www.spotify.com

[2] http://www.apple.com/music/

[3] http://about.twitter.com/de/company

[4] http://the.echonest.com

five clusters according to their musical features, we observe that on average, each user creates playlists within three different clusters and that 17% of all users create playlists in all five clusters, suggesting that users arrange different styles of music in different playlists. Complementary to that, we find that although nearly half of the users create playlists with classical and rap-style music, these playlists account only for 8 and 7% of all playlists. Moreover, we detect a cluster where 91% of all users create playlists in as it contains a form of "feel-good" popular music, serving as a common musical ground across all users. Our analyses also show that people do not necessarily group their music by genre. We consider the insights gained in this work to be useful for improved automatic playlist generation and music organization.

The remainder of this paper is structured as follows: In the following section, we present works related to the presented analyses. Section III subsequently introduces the data set and methods used to analyze user-created playlists. Section IV presents the results of the conducted analyses, which are further discussed in Section V where we also point out future work. Section VI concludes the paper.

II. Related Work

In literature, several studies about music organization can be found. Cunningham et al. conducted a study on how people organize CDs and MP3 files, based on interviews and on-site observations of focus groups. They found that facilities for creating playlists are a demanded feature [14]. In a later study, based on an online survey, Kamalzadeh found that people prefer a minimal amount of interaction. At the same time, users want the music to match their mood and want to be able to change the mood of the music played [1]. As for minimizing the required interaction with music systems, the automatic generation of playlists has been studied intensively starting from the early 2000s. We categorize these approaches into (i) approaches mainly utilizing content or metadata and (ii) hybrid approaches incorporating user feedback in addition to the data sources mentioned before.

With respect to (i), there are approaches that facilitate a seed song along with the traditional k-nearest neighbors approach to find similar songs to the given start song [10]. Further, approaches in which the user selects a start and an end song with a smooth transition in between [12] and approaches based on user-defined constraints [8] have also been proposed. The used constraints may be content-based, i.e., the tempo of a song, or based on meta-information like the genre [7], [8]. With respect to (ii), we find approaches incorporating the contexts-of-use. In this case, metadata of tracks is used to cluster similar songs to playlists and users were asked to judge the suitability of this cluster for certain contexts-of-use [9]. Besides this, also the skipping behavior combined with content-based features has been exploited. Skipping a song as an indicator for dislike is used in order to avoid adding songs with the same content-based features to the playlist as the skipped one [11].

Following up this prior research, in this work, we focus on how users facilitate their playlists on the music streaming platform Spotify. In contrast to [14] and [1], we approach this topic quantitatively using a broad user base gathered from the Spotify platform. This is done in order to lay a foundation for future music databases and libraries, recommender systems or new forms of playlist generation.

III. Data Set and Methods

In this section we provide details about our data set as well as the methods utilized for the performed analyses, before discussing and interpreting the results in the subsequent section.

A. Data Set

For the analyses presented in this work, we gathered a novel data set via the Spotify API. Our data set is based on the #nowplaying data set containing more than 56,817,896 listening events scraped from Twitter [3]. In order to get a initial list of users to crawl and complement this data set with user created playlists, we extract the usernames of users tweeting via Spotify. This way, we gather 1,137 Spotify users, organizing 796,024 distinct tracks in 18,296 playlists via the official Spotify API. In a second step, we enlarge the data set with content-based information crawled from the Echo Nest platform via their API[5]. As the Echo Nest and Spotify cooperate[6], we query the Echo Nest using the Spotify track identifiers. We are able to retrieve content-based features for more than 90% of the tracks. Tracks for which we could not retrieve content-based features were removed from the final data set, as our analyses require those features. The resulting data set contains 1,133 Spotify users, 18,146 playlists, 706,989 tracks and for each of the tracks, the acoustic features provided by the Echo Nest. On average, the data set features 18,25 (SD=19.07) playlists and 1,084.07 (SD=2,659.45) tracks per user. As for the acoustic features, we retrieve the audio summary of all tracks as provided by the Echo Nest. I.e., we extract the audio features danceability, energy, loudness, speechiness, acousticness, liveness and tempo. A detailed description of the extracted acoustic features can be found online[7].

B. Data Cleaning and Aggregation

As we aim to get a deeper understanding for music playlists, we have to filter for musical tracks within our data set. Thus, we restrict the data set to tracks with a *speechiness* of 0.66 or below. According to the Echo Nest documentation, tracks with a *speechiness* higher than 0.66 are most likely audio books or the like[7]. To analyze the acoustic features of each playlist, we aggregate the acoustic features of the individual tracks for each playlist in the data set using the arithmetic mean. To show the dispersion of the tracks forming a playlist, we state the mean as well as the mean absolute deviation (MAD) of each acoustic attribute in Table I. We make use of the MAD as it is a robust measure with respect to outliers [16]. The

[5] http://developer.echonest.com/docs/v4

[6] http://static.echonest.com/enspex/

[7] http://developer.echonest.com/acoustic-attributes.html

Attribute	MAD >Mean	%
tempo	0	0.00%
energy	61	0.34%
speechiness	39	0.21%
acousticness	1,392	7.67%
danceability	2	0.01%
loudness	18,145	99.99%
valence	101	0.56%
instrumentalness	978	5.39%

TABLE I: Aggregated Acoustic Features

table shows that except for loudness, the variance of each of the acoustic characteristics of the tracks inside a playlist is low and the MAD is rarely higher than the mean. Thus, we can conclude that aggregating the characteristics of the individual tracks to playlist characteristics using the mean is representative. Further, we argue that aggregating the loudness of the individual tracks to a playlist loudness is not reasonable: the variance among the loudness in the tracks of a playlist is too high. In 99.99% of all cases the MAD is higher than the mean. Therefore, we drop the loudness characteristic for the conducted playlist analyses.

C. Methods

In a first step, we aim to identify variables that explain most of the variance in the data set and hence, differences in the user-generated playlists in regards to acoustic features, which reflects RQ1. In order to find these variables, we conduct a Principal Component Analysis (PCA) [17]. After, the PCA is based on the standardized matrix to avoid problems with different scales. That is to say, we compute the Principal Components (PCs) using the correlations matrix in contrast to the covariance matrix. This is a common method for conducting PCAs [18]. In a second step, we make use of k-Means clustering [19] to aggregate playlists into groups (or types). We estimate k using the PCA conducted in the first step as proposed by Ding and He [20]. This clustering is done to answer RQ2 and hence, aims to find certain types of playlists. To find user types creating such playlists, we rely on several correlation and similarity measures as we aim to find correlations between users creating certain playlists in certain clusters.

IV. Results

In this section, we present the results of the analyses conducted using the methods described in Section III. Firstly, we elaborate the results regarding RQ1, finding groups of playlist, before focusing on the users and thus, on RQ2.

A. Groups of Playlists

Based on the aggregated data set described in the preceding section, we conduct a PCA. Figure 1 depicts a biplot of the first two Principal Components (PCs), where each playlist is represented as a dot. This allows to analyze half of the variation within the playlists data set. The first PC on the x-axis distinguishes *acoustic* and *instrumental* playlists from playlists focusing on *tempo* and *energy* as well as playlists focusing on *valence* and *danceability*. This is, as the loading vector of PC1 only has negative signs for *acousticness* and *instrumentalness* and thus contrasts those two attributes from the other attributes. By only using the first PC, we are able to explain 27% of the variation.

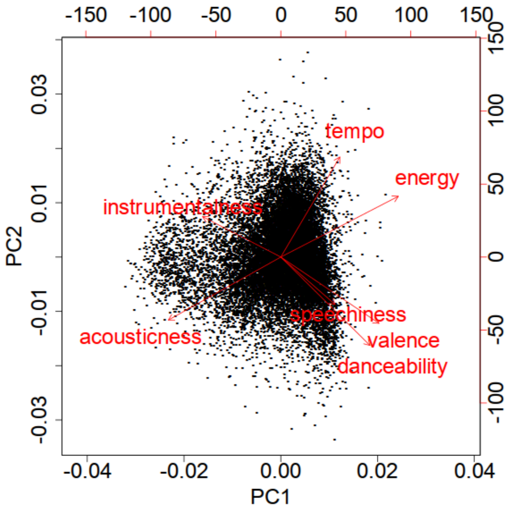
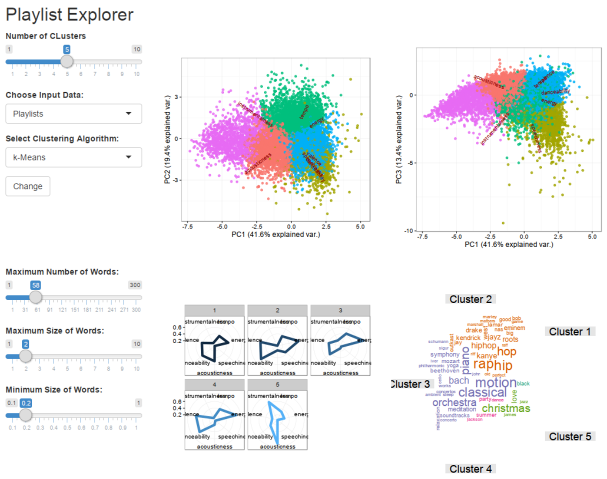


Fig. 1: Biplot using PC1 and PC2

Analogously, we observe that the second PC on the y-axis divides more *instrumental* playlists and playlists with high *tempo* and *energy* from playlists which are more *acoustic* as well as playlists with high *danceability*, *valence* and *speechiness* values. Again, this is as the loading vector of PC2 has negative signs for latter attributes, whereas the former three attributes are positively signed. By using the second PC, we are able to explain another 19% of the variation. By using our web-based analysis tool[8] as shown in Figure 2, we allow multimedia researchers to investigate arbitrary combinations of PCs. In this work, we complement our analysis by looking at PC3: PC3 separates tracks with high *speechiness* values from the rest. Using the first 3 PCs, we are able to explain 61% of the variance. Each further added PC adds 10% or less explained variance. Based on the findings of the conducted PCA, we aim to partition our set of playlists into clusters of playlists: *instrumental* and *acoustic* playlists, playlists focusing on *valence* and *danceability* along with *speechiness* and playlist focusing on *tempo* and *energy*. Hence, we apply k-Means clustering with $k = 3$ to $k = 7$. Clustering into 3 clusters leads to clusters that are based on the first two PCs (as described above), whereas clustering into 7 clusters leads to clusters based on each of to the 7 acoustical features. This is shown in Figure 3, where different k-Means solutions are plotted for different k. Each point represents a playlist, plotted against PC1 and PC2. The color and shape of the points represent the cluster membership. In order to formally determine the optimal number of clusters for our next analyses, we rely on the wide spread method utilizing the gap statistic

[8] http://dbis-pla.uibk.ac.at


Playlist Explorer

Number of CLusters: 5

Choose Input Data: Playlists

Select Clustering Algorithm: k-Means

Change

Maximum Number of Words: 58

Maximum Size of Words: 2

Minimum Size of Words: 0.2

Fig. 2: Web-Based Analysis Tool

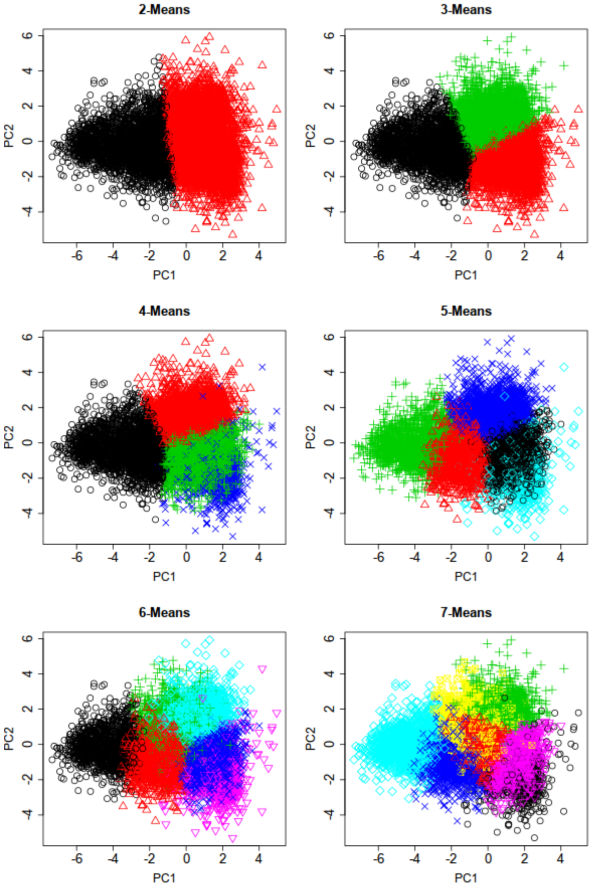


Fig. 3: k-Means for k between 2 and 7

[21]. This method is based on the “Elbow Curve” [22] or rather on the idea that it is important how much the within-cluster sum of squares (WCSS) decreases with an increasing number of clusters, as the WCSS naturally decreases with the number of clusters. In our approach, the gap statistic indicates that 5 clusters are an appropriate solution. The result is confirmed by plotting the “Elbow Curve”, which is in our case plotting the number of clusters vs. the WCSS. In a next step, we aim to get an overview of the acoustical attribute characteristics of the five clusters. Therefore, we visualize these as a radar diagram for each cluster as shown in Figure 4. This diagram shows the different features and their manifestation in the five clusters. Cluster 1 contains tracks focusing on *energy* and *tempo*, whereas cluster 2 contains tracks with high *speechiness*, *energy*, *valence* and *danceability* values. Cluster 3 is rather similar to cluster 2, besides the high *speechiness* values. This is, as the former one contains mostly rap music, in contradiction to the latter, which contains different forms of pop music. This observation is underpinned by the genre distribution as discussed in Section IV-B. Furthermore, we witness that high *danceability* values correlate with high *valence* values (Clusters 2 and 3). Cluster 5 contains tracks focusing on *acousticness* and *instrumentalness* as this cluster mostly contains classical music. Again this is reflected in the genre distribution.

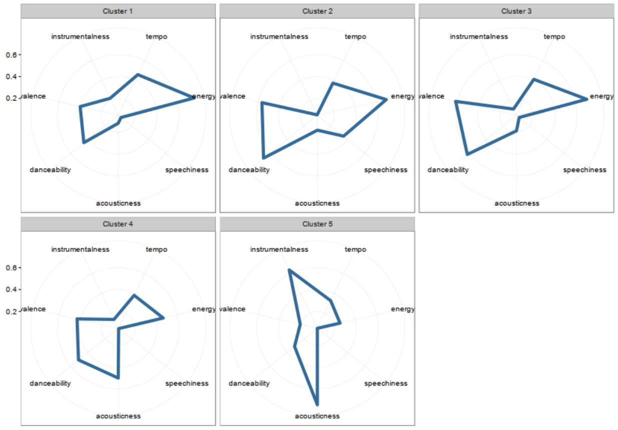


Fig. 4: Acoustical Characteristics of the Clusters

To answer RQ1, there exist differences based on the audio characteristics of playlists. By conducting a PCA we are able to explain 60% of the variance using the first 3 PCs: We observe, that the first PC separates *acoustic* and *instrumental* playlist from the rest. The second PC, separates playlists with high *valence* and *danceability* from the rest the third PC separates tracks with high *speechiness* values. Based on these characteristics, we are able to cluster playlists into 5 different groups using k-Means. This are already a valuable insights, however aiming to get a better understanding of the different clusters, we explore the genre distribution among each of the clusters in the next section.

B. Genre Distribution

In the following section, we provide a detailed analysis on the genres within the presented clusters.

We obtain genre information for each track using the genre tags provided by Spotify. To derive a genre distribution for each cluster we count the number of appearances of each genre in each cluster. In a next step, we look into whether there is a difference in the genre distribution among the clusters.

# Clusters	# Users	Relative Portion
≥ 1	1133	100.00%
≥ 2	923	81.47%
≥ 3	733	64.70%
≥ 4	478	42.19%
≥ 5	200	17.65%

TABLE II: User Distribution in Number of Clusters

Therefore, we rely on two traditional similarity measures (Jaccard [23] and Pearson Similarity [17]) to compute similarities between the different genres appearing in the individual clusters. Thus, in a first step, we count how many times each of the distinct genres occurs in each cluster. In a second step, we apply the two similarity measures on all pairs of clusters.

We observe one high correlation between clusters 1 and 3 ($r = 0.74$), which we lead back to the different forms of pop-music genres in those two clusters. Additionally, we observe a moderate similarity of several clusters. This implies that the same genres, mainly different forms of pop music, appear among several clusters. E.g., we can find the "popchristmas"-genre in all clusters. Hence, we argue that users do not necessarily group tracks by same ways as genres group tracks. In other words, users use the same genres in different playlists. In addition, we observe that the correlation coefficient is nearly 0 between Cluster 2 (the "rap Cluster") and Cluster 5 (the "classical music Cluster"), confirming that rap-style music is rather different from classical music. These results are consistent for Pearson and Jaccard Similarity.

Besides analyzing the genre distribution of the playlist-clusters, we also study the user distribution among the clusters in the next section.

C. Users among Clusters

In this section, we analyze the user distribution among the clusters representing playlists with similar acoustic features.

We investigate how many users create playlists only in a single cluster (i.e., they only listen to a single type of music in regard to acoustic features) and how many users create playlists in different clusters. In Table II, we state the number of users and the number of clusters they created playlists in. We observe that 64% of the users organize their music in playlists belonging to 3 or more clusters. About 17% of the users create playlists among all 5 clusters, the maximum. On average, a user is represented in 3.08 clusters with a median of 3 (SD=1.36). From the median and mean we can see that the number users with respect to the number of clusters is equally distributed. The average number of users per cluster is 631.60 with a median of 183 (SD=232.39).

We are also interested in whether we can find clusters, which are populated by the same users. I.e., whether if users create a playlist in cluster A, they are also likely to create a playlist in cluster B. Therefore, we look at the correlation between the clusters in terms of users having created playlists in those clusters. As the data can be considered ordinal or at least discrete between 1 and 54, which is the maximum number of playlists a user created within a cluster, we apply Spearman's rank correlation coefficient as shown in Table III.

	1	2	3	4	5
1	1.00	0.32	0.55	0.54	0.41
2	0.32	1.00	0.42	0.36	0.23
3	0.55	0.42	1.00	0.64	0.40
4	0.54	0.36	0.64	1.00	0.56
5	0.41	0.23	0.40	0.56	1.00

TABLE III: User-Cluster Correlations

We do not observe any strong correlation between the individual clusters ($\rho > 0.7$), nevertheless there are several moderate correlations ($\rho > 0.5$) between the clusters. It is worth to mention that cluster 2, the "rap cluster", does not have any moderate correlations with other clusters. Cluster number 5, the "classical music cluster" only shows low moderate correlation with cluster number 4. However, every cluster except cluster 2 (rap) does have these moderate correlations to cluster number 4, the "folk cluster", a cluster containing different forms of folk music according to the genre distribution. Further, clusters 1 and 3 also show a moderate correlation. With respect to the acoustic attributes of these two clusters, they are rather similar, except for the fact that cluster number 3, containing pop music, shows higher values for *valence* and *danceability*. We interpret this as "feel good music".

Complementary to this, to estimate the overall popularity of the clusters, we compute the number of users and playlists in each cluster as shown in Table IV.

Cluster	Users	%	Playlists	%	Pls./Users
1	768	68%	5,129	28%	6.68
2	427	38%	1,423	8%	3.33
3	1,032	91%	7,967	44%	7.72
4	793	70%	4,623	25%	5.83
5	447	39%	1,534	8%	3.43

TABLE IV: Users and Playlist per Cluster

We find that 91% of all users created playlists in cluster number 3, the "feel good music"–cluster. Also, 44% of all playlists are located in this cluster. Interestingly, nearly 40% of all users created playlists in the "rap" or "classical music" clusters, however playlists in those clusters only account for 7 and respectively 8% of all playlists. This means that high number of persons create playlists with rap or classical music, while at the same time, the number of playlists with respect to the total number of playlist is low. This means, that classical music or rap music can be considered as niche music with respect to the number of playlists but not with respect to the number of users.

V. Discussion and Future Work

Summing up our results, we find and explain differences in terms of acoustic features across the playlists using a linear PCA. Based on this, we cluster playlists into 5 clusters (or groups) of playlists using k-Means clustering. On average, a user is represented in 3 clusters (SD=1.36), which indicates that one user prefers different styles of music. This supports qualitative studies that people prefer different styles of music dependent on the intended use or the mood [1], [14] quantitatively and furthermore shows that those studies are also valid nowadays for music steaming platforms. Along with that we find, that the genre seems to be classifying music different to our classification based on acoustical attributes. As the same genres are present in several clusters and playlists, we can argue the classifying music for certain playlist using the genre is not the best way. Furthermore, we see that different types of music (in terms of acoustical attributes) are tagged with the same genre. Based on these findings we argue, that novel approaches for classifying tracks in music databases and libraries, as presented in the next paragraph, could be valuable to the users.

In future work, we plan to look into tagging each of the clusters with a certain moods or the intended use. As already mentioned in Section II, people want to have very little interaction with their music databases and libraries, but still want to get music matching their mood or their activities. Thus, a possible application could provide search facilities capable of finding music fitting a special situation. This is why tagging our clusters with this information would enable presenting music to users based on clusters matching their activities and moods. One approach will be to exploit the playlist names as proposed by [6]. Another possible application using our findings and data is to create user classifications, i.e. based on mining for association rules in our data set. Possible rules would be that users who create playlists in rap and classical music clusters are users creating playlists in all cluster or a users solely creates playlists on the rap cluster won't create a playlist in the classical music cluster.

VI. Conclusion

We presented an analysis of user-generated playlists on the music streaming platform Spotify. This is the first study to facilitate Spotify playlist data for a quantitative analysis. Our main contribution is an explanation of differences and commonalities among user created playlist. We show that "feel-good" popular music is serving as a common musical ground across all users. 91% of all users create at least one playlist in the "feel good music"-cluster. Additionally, we observed, that classical music and rap music can be considered as niche music with respect to the number of playlists, however not as niche music when considering the number of users. Furthermore, users creating playlists in both, the rap and the classical music cluster, are rare. Further, we found that users in general listen to different styles of music (or at least organize different styles of music in their libraries). Finally, in order to foster research in this field, the methods and the data presented are made available publicly and are aimed to be exploited for assisting users in navigating and organizing their tracks in music databases, libraries and on music streaming platforms. This should contribute to an increased usability of those applications.

References

[1] M. Kamalzadeh, D. Baur, and T. Mller, "A survey on music listening and management behaviours," in *Proc. of the 13th Intl. Symposium on Music Information Retrieval (ISMIR)*, 2012.

[2] D. Hauger, M. Schedl, A. Košir, and M. Tkalčič, "The Million Musical Tweets Dataset: What Can We Learn From Microblogs," in *Proc. of the 14th Intl. Society for Music Information Retrieval Conf. (ISMIR)*, 2013.

[3] E. Zangerle, M. Pichl, W. Gassler, and G. Specht, "#nowplaying music dataset: Extracting listening behavior from twitter," in *Proc. of the 1st ACM Intl. Workshop on Internet-Scale Multimedia Management*, 2014, pp. 21–26.

[4] E. Zangerle, W. Gassler, and G. Specht, "Exploiting twitter's collective knowledge for music recommendations," in *Proc. of the 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages*, 2012, pp. 14–17.

[5] M. Schedl and D. Schnitzer, "Location-aware music artist recommendation," in *Proc. of the 20th Intl. Conf. on MultiMedia Modeling (MMM)*, 2014.

[6] M. Pichl, E. Zangerle, and G. Specht, "Towards a context-aware music recommendation approach: What is hidden in the playlist name?" in *Proc. of the 15th IEEE Intl. Conf. on Data Mining Workshops (ICDM)*, 2015, pp. 1360–1365.

[7] M. Alghoniemy and A. H. Tewfik, "A network flow model for playlist generation," in *Proc. of the Intl. Conf. on Multimedia and Expo (ICME)*, 2001, pp. 329–332.

[8] J.-J. Aucouturier and F. Pachet, "Scaling up music playlist generation," in *Proc. of the Intl. Conf. on Multimedia and Expo (ICME)*, 2002, pp. 105–108.

[9] S. Pauws and B. Eggen, "PATS: Realization and user evaluation of an automatic playlist generator." in *Proc. of the 3rd Intl. Symposium on Music Information Retrieval (ISMIR)*, 2002.

[10] B. Logan, "Content-Based Playlist Generation: Exploratory Experiments." in *Proc. of the 3rd Intl. Symposium on Music Information Retrieval (ISMIR)*, 2002.

[11] E. Pampalk, T. Pohle, and G. Widmer, "Dynamic playlist generation based on skipping behavior," in *Proc. of the 6th Intl. Symposium on Music Information Retrieval (ISMIR)*, 2005, pp. 634–637.

[12] A. Flexer, D. Schnitzer, M. Gasser, and G. Widmer, "Playlist generation using start and end songs." in *Proc. of the 9th Intl. Symposium on Music Information Retrieval (ISMIR)*, 2008, pp. 173–178.

[13] M. Slaney and W. White, "Measuring playlist diversity for recommendation systems," in *Proc. of the 1st ACM workshop on Audio and music computing multimedia*, 2006, pp. 77–82.

[14] S. J. Cunningham, M. Jones, and S. Jones, "Organizing digital music for use: an examination of personal music collections," in *Proc. of the 5th Intl. Symposium on Music Information Retrieval (ISMIR)*, 2004.

[15] M. Pichl, E. Zangerle, and G. Specht, "Combining Spotify and Twitter Data for Generating a Recent and Public Dataset for Music Recommendation," in *Proc. of the 26nd Workshop Grundlagen von Datenbanken, Ritten, Italy*, 2014.

[16] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764 – 766, 2013.

[17] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine*, vol. 2, no. 6, pp. 559–572, 1901.

[18] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[19] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[20] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. of the Twenty-first Intl. Conf. on Machine Learning (ICML)*, 2004.

[21] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[23] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.