Multimed Tools Appl (2019) 78:2897–2920
https://doi.org/10.1007/s11042-018-5980-y

Predicting user demographics from music listening information

Thomas Krismayer[1] · **Markus Schedl**[2] · **Peter Knees**[3] · **Rick Rabiser**[1]

Received: 12 October 2017 / Revised: 25 January 2018 / Accepted: 9 April 2018 /
Published online: 7 May 2018

Abstract Online activities such as social networking, online shopping, and consuming multi-media create digital traces, which are often analyzed and used to improve user experience and increase revenue, e. g., through better-fitting recommendations and more targeted marketing. Analyses of digital traces typically aim to find user traits such as age, gender, and nationality to derive common preferences. We investigate to which extent the music listening habits of users of the social music platform Last.fm can be used to predict their age, gender, and nationality. We propose a feature modeling approach building on Term Frequency-Inverse Document Frequency (TF-IDF) for artist listening information and artist tags combined with additionally extracted features. We show that we can substantially outperform a baseline majority voting approach and can compete with existing approaches. Further, regarding prediction accuracy vs. available listening data we show that even one single listening event per user is enough to outperform the baseline in all prediction tasks. We also compare the performance of our algorithm for different user groups and discuss possible prediction errors and how to mitigate them. We conclude that personal information

✉ Thomas Krismayer
thomas.krismayer@jku.at

Markus Schedl
markus.schedl@jku.at

Peter Knees
peter.knees@tuwien.ac.at

Rick Rabiser
rick.rabiser@jku.at

[1] Christian Doppler Lab MEVSS, Institute for Software Systems Engineering, Johannes Kepler University Linz, Linz, Austria

[2] Department of Computational Perception, Johannes Kepler University Linz, Linz, Austria

[3] Institute of Information Systems Engineering, Faculty of Informatics, TU Wien, Vienna, Austria

can be derived from music listening information, which indeed can help better tailoring recommendations, as we illustrate with the use case of a music recommender system that can directly utilize the user attributes predicted by our algorithm to increase the quality of it's recommendations.

Keywords User trait prediction · Digital user traces · User demographics · Music listening habits

1 Introduction

Everyday online activities such as using social networks or microblog services or shopping and consuming media leave digital traces, that indicate products or topics the user is interested in. These traces are recorded and many services use systems to recommend new items based on items the user selected or rated in the past (e.g., the social music platform Last.fm or the online movie streaming service Netflix) [15].

It has been shown that many of the digital traces that are left by the users can also be exploited to predict additional information about them such as predicting a person's location from their tweets [2] or predicting personality traits from Facebook Likes [18]. In this work, we focus on digital traces on the social music platform Last.fm, and use various different sources of information either directly from data available via the Last.fm API or extracted from the collected data to infer personal information of users, such as age, gender, and nationality.

We consider this a highly relevant topic with respect to digital media consumption and social media usage behavior for two reasons: on one hand, gaining a better understanding of the users will help in better understanding the contents of the media they are using, and thus help in creating more "semantic" indexing techniques also better supporting users finding what they need. On the other hand, we are interested in how much this seemingly "harmless" and therefore often naïvely and thoughtlessly shared information can be used to derive additional information about the users. This second aspect exhibits direct ties to concerns regarding privacy and profiling.

We build on our previous work on this topic – presented in [19]. In this work, we report insights that go beyond the prior findings and are based on additionally analyzing the obtained results in detail. Specifically, as the success of such a prediction system depends on its performance and accuracy, we analyze the possible negative influences on our algorithm in detail, particularly the influence of the available user data. This allows us to better understand the flaws of our prediction pipeline – especially its dependence on user distribution and similarity – and as a basis for further post-processing steps to mitigate these flaws. We also extended the discussion of related work and provide more details on the dataset we used. Additionally, we describe the use case of a music recommender system that can directly utilize the user attributes predicted by our algorithm to increase the quality of it's recommendations.

More specifically, we investigate the following research questions:

- *(RQ1) To which extent is it possible to predict the age, gender, and nationality of the users based on their listening events and related information* (such as how the listening behavior changed over time)?
- *(RQ2) In which way does prediction accuracy depend on the available user data* (i. e., the number of listening events)?

- *(RQ3) What kind of errors are made by our prediction pipeline and how can they be explained and mitigated?*
- *(RQ4) How can the user attributes predicted by our algorithm be used to increase the quality of a music recommender system?*

The results of our proposed approach can be utilized to enrich the input for recommender systems (e. g., to replace missing values for collaborative filtering approaches) or directly for recommending new items (e. g., artists that are popular in the country or within the age group of the user). In further steps the system could also be used to directly predict topics (e. g., genres) or items (e. g., artists or songs) the user is interested in, thus improving the user experience.

The remainder of this paper is structured as follows. In Section 2, we discuss literature related to the prediction of user traits from digital traces. Section 3 provides a detailed description of the dataset used in our experiments. We introduce the actual algorithm for predicting user traits in Section 4. In Section 5, we describe the experiments performed and the results gained, which are then analyzed in detail in Section 6, especially with regard to possible errors and how to mitigate them. A short example use case that illustrates how the predictions produced by our system can be utilized to improve music recommendations is then given in Section 7. Finally, in Section 8 we wrap up the paper with a conclusion and an outlook on future work.

2 Related work

In this section, we discuss work on automated prediction of user traits from digital traces, structured according to the source of collected user traces. Additionally we present work on applications utilizing these user traits, e. g., to improve the performance of recommender systems.

2.1 Prediction of user traits

Kosinski et al. [18] show that user traits can be predicted based on the **Facebook** Likes of a person. The predicted values include basic profile information, such as age and gender, but also highly personal attributes, such as sexual orientation, ethnicity, political views, and personality traits. The prediction is based on the Likes of 58,000 Facebook users, for whom demographic profiles and psychometric tests are available. A follow-up study, conducted by Youyou et al. [39], shows that personality judgments made from Facebook Likes can be even more accurate than those of close friends or family members. Golbeck et al. [9] show that the personality of Facebook users can even be predicted based only on their publicly available profile information. Finally, Ortigosa et al. [23] predict the personality traits of Facebook users from indicators such as the number of friends and the number of posts per month.

The algorithm described by Cheng et al. [2] estimates the location of **Twitter** users based on the text of their tweets. The estimation is entirely content-based and does not rely on meta-data, such as profile or network information. The proposed algorithm is trained on Twitter users in continental USA whose locations are known and then predicts the user location by inferring probabilities for cities from the microblogs. In their experiment, Cheng et al. report that 51% of the users were placed within 100 miles of their actual hometown. Conover et al. [6] show in their work that using content analysis (e. g., analyzing

the text of Tweets) or network analysis (e. g., constructed from Retweets) can be used to predict the political alignment of Twitter users. In [36] Volkova et al. predict perceived psycho-demographic attributes – including age and gender – for Twitter users based on their interests. It is shown that for some user attributes the user interest (extracted from the accounts the users follow) allows to extract results very similar to state-of-the-art content-based approaches (that utilize Tweet texts). This approach can still be used if it is not possible to collect texts created by the users (e. g., due to account settings).

Predictions of user traits can also be made from usage patterns of **mobile phones**: In [7], de Montjoye et al. use mobile phone logs to predict the personality of the phone owner. The indicators calculated from the logs include the number of calls, time to answer a text, calling routines, and number of interactions per contact. In the work of Malmi et al. [22] six different binary user traits (including age and gender) are predicted based on the apps the smart phone users install. Additionally they analyze which apps give the most information for the prediction tasks and show that, e. g., Snapchat is highly indicative for young users (32 years and younger) or that period-tracking apps are helpful for predicting gender.

Most closely related to our paper is work that exploits **Last.fm** data to predict listener characteristics. Liu et al. [21] estimate the gender of Last.fm users based on their listening history. Additionally, the age is estimated in a binary form as under or above 24 years. The features for the classification are constructed purely from the listening events of the user and are based on three factors: the listening timestamps, the meta-data of the song and the artist (e. g., artist and song tags), as well as signal features of the songs. For both tasks, a support vector machine classifier (SVM) with RBF kernel is used and the average of five runs with 80% of the users as training set is reported. The accuracy for age is 71.1%; the accuracy for gender is 66.1%.

The approach described in the work by Wu et al. [38] estimates gender and age of Last.fm users based on music meta-data. Their algorithm uses the songs that the user most frequently listens to. In contrast to Liu et al., the approach does not exploit temporal information, nor any audio-based features. The authors describe two different ways to generate features for the user: Term Frequency – Inverse Document Frequency (TF-IDF) combined with Latent Semantic Indexing (LSI) and Gaussian Super Vectors (GSV). For both tasks, SVM with RBF kernels are used in a two-fold cross validation. The reported accuracy for gender estimation is 78.87% and 78.21% for GSV and TF-IDF, respectively. For age estimation a mean absolute error of 3.69 and 4.25 is reported for the GSV and the TF-IDF approach, respectively.

In contrast to these two existing works [21, 38], our main ***contributions*** are: (i) we present a novel approach for the prediction of user traits from music listening habits that combines multiple sources of information and uses PCA-compressed TF-IDF-like features, (ii) we also support the prediction of user nationality, (iii) we conducted experiments with users with a very limited number of listening events to assess performance in cold-start situations; (iv) we compare different machine learning classification and regression algorithms; (v) we analyzed the predictions of our approach in detail; and (vi) we present an example application of the user attributes predicted by our algorithm.

2.2 Application of user traits

In [30] and subsequently in [31] it is argued that music recommender systems often fail to produce satisfiable results, because they do not use all the information needed to find the music that perfectly fits the current situation, i. e., a user's mood or location. Specifically,

music perception is influenced by four different factors: music content (e. g., rhythm), music context (e. g., semantic labels), user context (e. g., mood) and user properties (e. g., music preference, age, gender, etc.). To produce better fitting music recommendations all of these factors should be considered by a recommender system. The user traits that are predicted by the algorithm described in this paper – which are part of the user properties – can therefore be used to improve music recommendations, if these pieces of information are not (or only party) available (e. g., to compensate missing profile information).

In [8] Fuller et al. categorized users of music streaming services into seven distinct groups (personas) depending on the way they use and interact with the streaming services. Examples for these personas are "Active Curators", who often generate and listen to playlists and "Music Epicureans", who are likely to listen to an entire album. The authors also analyzed the similarity between the different personas and additionally found that users with different personas prefer different music streaming services – e. g., Active Curators prefer Spotify, while Music Epicureans prefer YouTube.

Zuang et al. [40] combine diverse sources of information about Flickr users – such as their images, tags, friend list and comments – to estimate the social strength between users. Combining these heterogeneous data the authors construct a social strength graph using a kernel-based learning approach. The authors show that this graph can be used for various different problems such as recommending friends or user groups.

Cheng et al. [1, 3] presented a music recommender system that utilizes the user's location-related context in combination with global music popularity trends. The music popularity trends are extracted by collecting tweets with the hashtag #nowplaying or #np, using the Twitter API. This information, in combination with information on the music content, outperforms other recommender systems at different locations.

In [4] Cheng et al. present a text-based music recommendation system that provides recommendations that fit both the search term given and the music preferences of the user. The preferences of the user are estimated based on the audio content and the tags associated with songs the user frequently listened to. Similarly, in [5] the authors present a text-based music retrieval system that utilizes user traits to improve music predictions. However, instead of capturing and using the music preferences of the user, the system captures music preferences of different age groups and genders. These preferences are then used to re-rank the recommended items based on the age and gender of the current user. This approach illustrates another way the user traits predicted by our system could be used, i. e., to predict the age and gender for users that did not enter this information in their profile. It is also possible to use country-specific preferences to improve the recommendations.

3 Dataset

The dataset used in our experiments is a subset of the LFM-1b dataset[1] published in [28]. The LFM-1b dataset was created using the Last.fm API, which allows the collection of users' profile information (including age, gender, and country) as well as listening events for these users, where a listening event is defined by user, artist, album, track identifiers, and a timestamp. The time period of included listening events is January 2005 to August 2014.

[1] http://www.cp.jku.at/datasets/LFM-1b

The data for the LFM-1b set was acquired by first fetching the overall 250 top tags,[2] to, in turn, gather their top artists.[3] For these artists, the top fans[4] were retrieved, which resulted in 465,000 active users. Subsequently, the listening histories of a randomly chosen subset of 120,322 users[5] were obtained. Please note that some of the Last.fm API endpoints are deprecated.

The LFM-1b dataset additionally includes scores describing the listening behavior of the users. These scores include novelty, i. e., percentage of new artists in a specific time period, "mainstreaminess", i. e., how well the preferences of the user fit to the average preferences of all users, and different listening counts, e. g., the absolute number of distinct artists the user listened to, the average number of events per week, and the relative number of events for one specific day of the week.

3.1 Dataset selection and characteristics

Discarding from the LFM-1b dataset users with missing demographic information or less than 500 listening events, 12,181 users remained for our experiments. For these users a total of 6,736,824 listening events – including 818,814 individual artists – are contained in the original dataset and are used in our experiments. The resulting subset of the LFM-1b dataset is available for download,[6] which contains LFM-1b user identifiers and demographics.

This allows to use the same dataset for all three prediction tasks (age, gender, and country). The restriction to users with at least 500 listening events ensures that all users have the same number of listening events for the experiments with listening event subsets (see Section 3.3).

In addition to the information provided in the core LFM-1b dataset, we extracted weighted artist tags for the artists in the LFM-1b subset used in our experiments, again exploiting the Last.fm API.[7] Examples for such tags include genre information (e. g., "rock") and music characteristics (e. g., "female vocalist"). Additionally, these tags contain a weight from 0 to 100. We use these weighted tags later in our experiments for the prediction of the user traits and to identify artists that produce similar music.

The dataset used in our experiments eventually contains users from 144 countries with 72.5% of them being male and the average age being 25.6 years. In terms of number of users, the top countries in our dataset are: USA (19% of all users), Russia (8.9%), Germany (8.4%), Brazil (7.9%), Poland (7.8%), Great Britain (7.8%), and the Netherlands (2.6%). This distribution is similar to the distribution among the users in the entire LFM-1b dataset [29].

3.2 Balanced gender dataset

Due to the high share of male users in the dataset the baseline for the accuracy of gender prediction is rather high (72.5%). Although the best classifiers perform significantly better

[2] http://www.last.fm/api/show/tag.getTopTags

[3] http://www.last.fm/api/show/tag.getTopArtists

[4] http://www.last.fm/api/show/artist.getTopFans

[5] http://www.last.fm/api/show/user.getRecentTracks

[6] http://www.cp.jku.at/datasets/LFM-1b/LFM-1b_subset_MTAP2018.csv

[7] http://www.last.fm/api/show/artist.getTopTags

(81.4%, cf. Section 5.4), it is difficult to assess the performance of these classifiers. To overcome this problem when investigating the first research question for gender, during all experiments for gender prediction, we created multiple datasets, for which the users are filtered by selecting all female users and randomly selecting exactly as many male users. The datasets resulting from this procedure contain a total of 6,698 users (compared to the 12,181 users of the entire dataset) with a 50% share of female users.

3.3 Sampling listening event subsets

We sampled small random subsets from the listening histories of users with 1, 2, 5, 10, 20, 50, 100, 200, and 500 listening events per user to investigate to what degree the accuracy of predictions depends on the number of listening events used for training. The prediction results gained from these subsets build the basis for our evaluation of RQ2, i. e. to analyze predictions based on very limited data available for each user.

3.4 Age groups

For the feature selection and the analysis of results for age prediction the users are split into eight distinct age groups also used in [32]. These groups contain the users in the age intervals [6–17], [18–21], [22–25], [26–30], [31–40], [41–50], [51–60], and [61–100]. Figure 1 shows the distribution of users in these age groups.

4 Prediction of user traits

For prediction of user traits, we developed three models, one for age, gender, and country, respectively. Each model is built individually and does not use results from the other models. Furthermore, the models are built entirely from the listening data of the users, meta-data of the artists, and extracted user information. Therefore, e. g., for the prediction of age, the model does not use the gender or the country of the user.

4.1 Experimental setup

The prediction models are evaluated with a 10-fold cross-validation on the dataset introduced in Section 3. All steps for the prediction pipeline (feature selection, feature

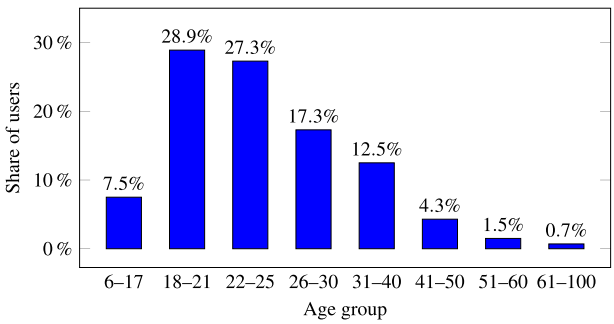

Fig. 1 Share of users per age group

vector generation, dimensionality reduction, classification/regression) were individually performed for the different user traits age, gender, and country. The calculations for all steps are based solely on the training set; this also implies that the selected features and the dimensionality reduction rules are different for each fold of the cross-validation.

4.2 Feature selection

We construct an individual feature vector for each user which contains elements from three separate sources – the first part is based on artist listening information, the second part on artist tag information, and the third part on additional user information provided as part of the LFM-1b dataset. These three parts are created independently from each other. The first two parts are vector normalized separately, for the third part this cannot be done because of the different ranges for the individual scores as we will explain below. Finally, the three parts are merged to create one feature vector per user ("early fusion").

The *first part* of a user's feature vector (*artist listening information*) is created by selecting 10,000 artists based on the number of users that listened to them. The first 5,000 artists that are selected are the artists that have the most different users in the overall training set that listened to them at least once. This also means that the first half of artists is selected independently of the task. The other 5,000 artists are selected based on their number of different listeners in user-groups chosen for the specific task. This means the users in the training set are split into distinct groups and the artists with the most users listening to them for each of the groups are selected. Artists that were selected during the first step or for another user group cannot be selected a second time.

The selection of artists for the prediction of age is based on the age groups introduced in Section 3.4, which means the most popular artists for each age group are selected. For the gender prediction the algorithm chooses the artists with the most male and female listeners, respectively. Finally for the country prediction task the groups comprise the countries with the most users in the training set. The dataset contains 144 different countries, however the feature selection only takes into account the 25 most common countries within the training data to concentrate on the most crucial user groups. For the whole dataset the 25 most common countries contain 88.5% of all users.

The *second part* of a user's feature vector (*artist tag information*) is created by selecting 10,000 tags in the same way as the artists for the first part of the vector. The tags with the most users that listened at least once to an artist associated with this tag (with a tag weight higher than 0) are selected. The first 5,000 tags are selected based on the overall training set, while the second half is selected based on the same user groups as for the artists.

The *third part* of the feature vector contains 42 *additional scores* for each user, comprising scores for novelty (i. e., how many new artists did the user listen to in a given time period), mainstreamness (i. e., how well do the genre preferences of the user fit to the overall genre preferences of all users in the dataset), and various listening event counts (e. g., the average number of listening events per week). These scores are part of the dataset used and are available for each user.

The differences in the range of the scores makes a vector normalization of the third part pointless. For instance, the novelty scores of a user are calculated in the interval [0, 1], while the count values of listening events have no boundary and are above 10,000 for some of the users.

4.3 Feature vector generation

The entries for the first two parts of the feature vector of a user are calculated in the form of TF-IDF values for a term t (i. e., an artist or a tag) and a document d (i. e., the listening history of this user) as:

$$\text{tf-idf}(t, d) = (1 + \log(f_{dt})) \cdot \log\left(\frac{n}{f_t}\right) \tag{1}$$

where n is the number of users in the training set, and f_t is the number of users with at least one listening event containing t, i. e. an event with this artist or with an artist that is labeled with this tag (with a tag weight higher than 0).

For artists f_{dt} is simply the number of listening events with the artists, however for tags the value also takes the tag weight into account:

$$f_{dt} = \sum_{e \in E} weight(a_e, t') \tag{2}$$

where E is the listening history of the user, a_e is the artist of listening event e, and $weight(a_e, t')$ is the tag weight for tag t' and artist a_e, which is 0, if the artist is not connected to t'.

4.4 Dimensionality reduction

The feature vectors that result from the previous step have a total of 20,042 dimensions, which are reduced in a dimensionality reduction step using Principal Component Analysis (PCA) [14]. The PCA is performed on the combined first two parts of the feature vector (i. e., 20,000 dimensions) to ensure that correlations between artist and tag features can be resolved.

The 42 dimensions in the third part of the vector are not transformed because of their difference in range (cf. Section 4.2). This difference would cause the reduction algorithm to overly concentrate on these feature that contain a much higher variance compared to the other features.

In this step the number of features is reduced from 20,000 to 450 – plus the 42 unaffected features. The new number of features results from adding 50 features as long as the variance gained stays above 1.5% for the 50 new features – this results in 450 features for each of the three tasks. The dimensionality reduction is performed in Python using the library scikit-learn [24] and the transformation is calculated based solely on the training set. The compressed feature vectors for the test set are then constructed using the same transformation rules.

4.5 Predictions for listening event subsets

For the predictions based on listening event subsets (cf. Section 3.3) only the PCA-compressed first and second part of the feature vectors is used. The third part of the vector includes information that is not available in a cold-start-like situation that is simulated with these experiments and can therefore not be used. For instance, the novelty score represents an indicator of how the listening behavior of the user changes over time – an information that the system cannot estimate for a user, who just has one single listening event.

The classification/regression algorithm is trained on the original user vectors containing all listening events for the users in the training set. Based on this model the predictions for

all subsets of the test set are made. Due to the vector normalization the algorithm is able to deal with the different number of listening events and artist tags.

5 Experiments and results

To elaborate on RQ1 – i. e. to which extent it is possible to predict age, gender, and nationality of users – we build different supervised models based on the reduced feature vectors resulting from the dimensionality reduction. The models are constructed using a selection of diverse machine learning classifiers and regressors. For this purpose, we use the Java API of the open source library Weka [11]. In this section, we present the results for the individual experiments using the same evaluation methods as in [21, 38] (i. e., mean absolute error for age and accuracy for gender). Our experiments additionally contain predictions for nationality, which are also evaluated regarding accuracy.

We also evaluate the performance of the best classifiers on the reduced listening event subsets – thereby addressing RQ2, which deals with the predictions based on limited data – and the datasets with balanced gender share. We compare the results for all tasks to a baseline (detailed below) to help interpret their quality.

5.1 Learning algorithms

For the prediction of the results a variety of different supervised classification and regression techniques are used. The following algorithms achieved the best results:

Support Vector Machines (SVMs) separate two classes by defining a border function in a potentially higher dimensional space such that data points from the two classes lie on the different sides of the border. SVMs can also be used in regression tasks by creating a function such that all data points fall within a given maximum error margin. In both cases the class or value for a new data point is then predicted with the generated (border-)function. For our experiments the predictions are made using implementations of the Sequential Minimal Optimization algorithm (SMO [12, 17, 25] and SMOreg [33, 34]).

M5P [26, 37] is a decision tree algorithm enhanced with linear regression, which can be used as a decision criterion for some of the nodes within the tree. Based on this algorithm, **M5Rules** [13, 26, 37] creates a decision list that is filled with rules from decision trees built with M5P.

Linear regression generates a regression function as a linear combination of the features. Similarly **logistic regression** [20] predicts the class of a data point based on a linear combination of the features. We use the two logistic regression algorithms Bayesian Logistic Regression and Simple Logistic.

Naïve Bayes [16] and **DMNBtext** [35] use Bayes' theorem to predict the class of a new instance based on the probabilities for the different classes inferred from the training instances.

5.2 Baseline

The baseline for the given tasks represents a trivial lower bound for the results of the classifiers. For the classification tasks, the baseline used is a classifier that predicts the majority

Table 1 Mean absolute error for age prediction (best results)

Classifier	Settings	Mean absolute error
SMOreg	RBF Kernel	4.13
SMOreg	Normalized poly Kernel	4.17
SMOreg	Poly kernel	4.20
Linear regression		4.36
M5P		4.40
M5Rules		4.40
SMOreg	PUK	4.71
ZeroR		6.23

class of the training set for all instances of the test set. E. g., for country prediction the baseline is a classifier that predicts the country with the most users in the training set for all users in the test set. In case of a regression task, the classifier predicts the average value in the training set for all instances of the test set – e. g., the average age of the users in the training set. For both cases the calculation is done with Weka's ZeroR classifier [11].

5.3 Age prediction

Table 1 shows the algorithms that achieved the lowest mean absolute error for predicting the age of the users. The support vector regression (SMOreg) outperforms all other algorithms with three of the four kernels available for this task. The lowest error (4.1; achieved with the RBF kernel) is 66.3% of the error achieved with the baseline algorithm. The Linear Regression achieves a slightly better result than M5P and M5Rules. The baseline for this task is 6.2 (calculated with ZeroR).

The results for the age prediction based on the subsets of limited listening events can be seen in Fig. 2. We achieved these results with the SMOreg algorithm using the RBF kernel, which produced the best results for the entire dataset. These experiments simulate a cold-start problem, where the system is presented with almost no information about the user. Therefore the algorithm is only given the information derived from the listening events (i. e., the artist and tag information in the first and second part of the feature vector; c. f.

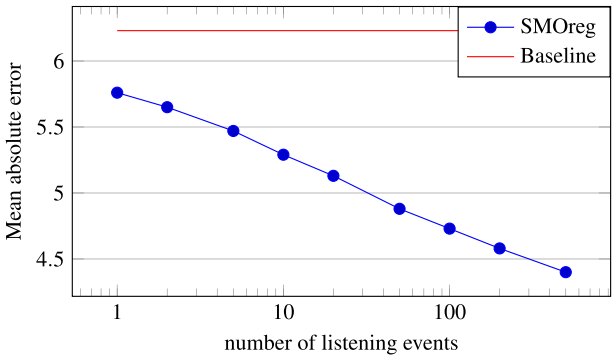

Fig. 2 Error for age prediction on listening event subsets

Table 2 Accuracy for gender prediction (best results)

Classifier	Settings	Accuracy
Bayesian logistic regression	Gaussian prior	81.36%
SMO	Poly kernel	81.24%
Simple logistic		80.43%
SMO	Normalized poly Kernel	78.06%
SMO	RBF Kernel	78.33%
DMNBtext		77.22%
SMO	PUK	76.31%
ZeroR		72.51%

Section 4.2) and not the 42 additional scores in the third part of the feature vector, which e. g., contain information on how the user preference changes over time.

Just one single listening event is sufficient to predict the age of the user more accurately than the baseline approach (5.8 vs. 6.2). The error of the prediction decreases steadily with an increasing number of listening events. Also, the final prediction that uses all of the available listening events achieves an error even lower than the prediction based on 500 listening events per user.

5.4 Gender prediction

The baseline for the gender prediction is 72.5%. As a result of each of the training folds having a majority of male users, this is the share of male users among the dataset (see Section 3). Table 2 shows the performance of the best classifiers for this task. The algorithm achieving the best results is the Bayesian Logistic Regression. This algorithm, which was developed for text categorization, benefits from the features of the feature vectors including clustered TF-IDF values, because TF-IDF weighting is an approach developed as basis for text analysis and text categorization. Both the support vector classifier (SMO) and the logistic regression algorithm (Simple Logistic) achieve results very close to the Bayesian Logistic Regression. The other algorithms yield far lower accuracy.

5.4.1 Balanced gender dataset

To compensate for the uneven gender distribution in the dataset, datasets with uniform gender distributions have been created, as detailed in Section 3.2. In order to ensure that the experiments on this dataset are not influenced by the listeners that are randomly picked for classification, the filtering is performed five times, the experiments are performed on each of the resulting datasets, and results are reported averaged over the five runs.

Due to the resampling of the dataset to achieve equal distribution of gender, the baseline for this task is obviously 50%. The results for the three classifiers that performed best on the whole dataset are given in Table 3. We report the average and the standard deviation over the five runs – to assess how representative the average results are. As it can be seen the standard deviation is very low for all three classifiers – the highest value being 0.36% for the Bayesian Logistic Regression – which shows that the influence of the users selected for each of the experiments on the performance of the classifiers is very low. It can also be seen that all three classifiers perform between 4.2% (SMO) and 4.5% (Simple Logistic) worse

Table 3 Accuracy for gender prediction on the balanced dataset (best results)

Classifier	Settings	Accuracy
SMO	Poly kernel	77.06% ± 0.24%
Bayesian logistic reg.	Gaussian prior	76.91% ± 0.36%
Simple logistic		75.88% ± 0.33%
Baseline		50%

than the same classifiers trained on the whole dataset (cf. Table 2), but have to be compared to a much lower baseline. The accuracy for the SMO using a poly kernel is 154.0% relative to the new baseline; for the complete dataset the Bayesian Logistic Regression achieves a relative accuracy of only 112.2% compared to the baseline.

The average results for the five runs of the balanced gender subsets using the Bayesian Logistic Regression and the SMO can be seen in Fig. 3. Both classifiers achieve very similar results for all listening event subsets and are able to achieve results better than the baseline with just one single listening event (up to 54.5% with the SMO classifier). The results improve steadily with additional listening events and also improve from 500 listening events to the overall result. For the experiments using subsets of the listening history of the users, only the features based on artist and tag information (based on the respective subset) are used. The additional user features given in the dataset cannot be used, as they contain information that is not available in a cold-start situation, which is simulated with these experiments.

5.5 Country prediction

Our third task is the prediction of the listeners' nationality. The baseline for this task is 19.0%, which equals the share of the most common country (USA) in the dataset. The classifiers that achieve the best results can be seen in Table 4. The two classifiers that perform best are the logistic regression algorithm (Simple Logistic) and the support vector classifier (SMO) which achieve 69.4% accuracy. The accuracy of the Simple Logistic algorithm is more than 3.6 times as high as the baseline.

The results for the reduced listening events can be seen in Fig. 4, which include the results for the two best performing classifiers for the test set with all events (cf. Table 4). Similar to the predictions for age and for the balanced gender sets, both classifiers are able to beat

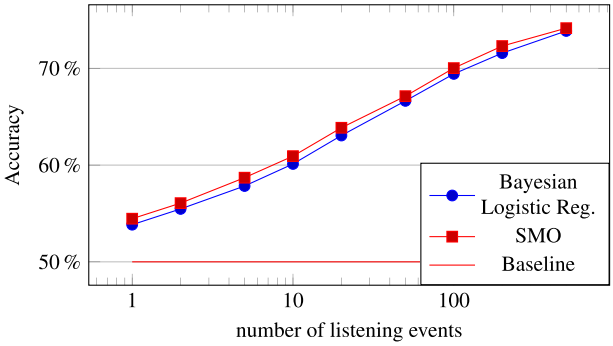

Fig. 3 Accuracy for balanced gender prediction on listening event subsets

Table 4 Accuracy for prediction of countries (best results)

Classifier	Settings	Accuracy
Simple logistic		69.37%
SMO	Poly kernel	69.36%
DMNBtext		63.11%
SMO	RBF kernel	59.97%
SMO	Normalized poly kernel	59.57%
Naïve Bayes		57.39%
ZeroR		19.03%

the baseline with just one single listening event (22.2% accuracy for the SMO) and improve steadily with additional listening events. As described in Section 4.2, these experiments use only the features generated from artist and tag information to simulate a cold-start-like situation.

5.6 Comparison of results with existing work

In the related work (cf. Section 2), the works of Liu et al. [21] and Wu et al. [38] have been introduced, which also target the prediction of user traits from music listening data.

The authors of [21] use the publicly available Last.fm 1K-users dataset to predict the gender and age of the users. This set contains users, for which user traits are missing. For the two experiments, the users, for which the respective trait is missing, are removed from the dataset. All the experiments are evaluated performing five runs with 80% of the users as training set and reporting the average of the results.

For this experiment we evaluated our approach with five-fold cross-validation, which also represents the average of five runs with 80% of the users as training set and additionally ensures that every user is part of the test set exactly once. Additional user information as in the Last.fm-1b dataset is not given and could therefore not be used.

For the gender prediction male users are removed from the dataset in order to create a set with a 50% share of female users. To lower the influence of the selected male users on the result we performed five runs of five-fold cross-validation – selecting different male users for each run – and reported the average result. The result achieved by our system is

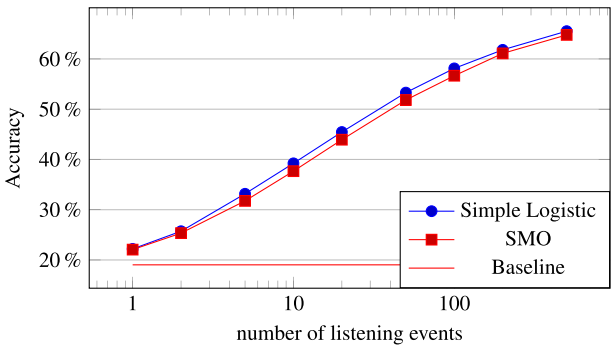

Fig. 4 Accuracy for country prediction on listening event subsets

72.9% (using Bayesian Logistic Regression with Gaussian prior), compared to an accuracy of 66.1%, which is the best result any approach in [21] achieved.

For the age prediction the authors split the user into the two classes "adolescents" (24 years and younger) and "adults" (25 years and older). The best result achieved by [21] is 71.1%, compared to 72.4% achieved by our system (with Bayesian Logistic Regression using Laplace prior).

The authors of [38] use their own dataset to predict the age and gender of Last.fm users. Therefore it is unfortunately not possible to test our approach on their dataset; also the different number of users (96,807 vs. 12,181 users) and the distribution of users (e. g., 66.2% vs. 72.5% male users) make a direct comparison of the received results pointless.

6 Analysis of results

In this section we introduce a similarity metric for user groups and analyze the results produced by our algorithm in detail. We compare the performance of the algorithm on different user groups and show that some of the errors are linked, e. g., to the user distribution and the similarity between specific user groups in the dataset. This analysis is performed to address RQ3, i. e., to explain and mitigate errors made during the user trait prediction.

6.1 User group similarity

To explain some of the relatively good or bad prediction results, we calculate the similarity values between selected user groups in the dataset. A relatively high similarity between two user groups can explain a low accuracy for one or both of the groups, as the classifier might wrongly assign users from one of the groups to the other.

To calculate the similarity between two groups of users we generate a genre vector for each user, sum up the vectors of the users within each group and calculate the cosine similarity between the two vectors. Each entry of a genre vector represents one genre in the list provided by Freebase [10]. From the 1339 genres that exist as tags in our dataset – which are 67% of the genres listed by Freebase – a feature vector x is generated for each user. The values x_t of the vector are calculated as:

$$x_t = \Sigma_{a \in A} count(a) * weight(a, t) \tag{3}$$

with t being a tag identified as genre, A being the set of artists labeled with this tag in the listening history of the current user, $count(a)$ being the number of listening events for this artist and the current user, and $weight(a, t)$ being the corresponding tag weight. All feature vectors are vector normalized and the final vector for the group is calculated as the sum of these vectors. We then calculate the final similarity value for two groups as the cosine similarity of their vectors.

For the analysis of the country predictions, the similarity between the users from different countries will be used as one way to explain the error between these countries. We assume users from countries with very similar genre preferences might be more difficult to distinguish than users from countries with rather diverse genre preferences.

6.2 Age

For the analysis of the age predictions we examined the mean absolute error for each of the age groups (cf. Section 3.4) – shown in Fig. 5. The results show that the error is smaller for

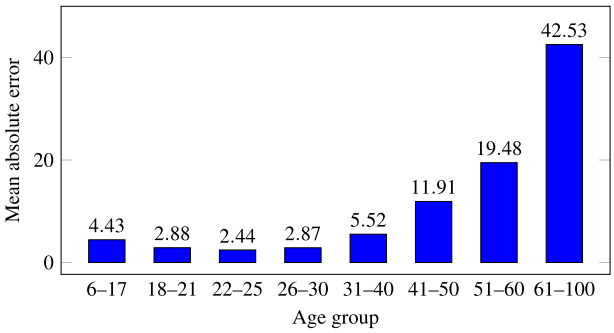

Fig. 5 Mean absolute error for age groups using SMOreg

age groups with a large number of users (see Fig. 1) and for groups closer to the average age (25.6 years) and median age (24 years). Especially for the groups containing the users with an age of 41 years and above the error increases dramatically. These groups have the highest distance to the average and median age and additionally contain the lowest amount of users.

Table 5 shows for each of the age groups how their users were classified. The cells containing the largest share of users from each age group (i. e., of each row of the table) are in bold. For this table the last three age groups are combined to one group. This is done because these groups contain a very low amount of users (see Fig. 1) – the new combined group [41–100] is still smaller than any of the other groups. Additionally, the regression algorithm does not assign an age of 51 or above (i. e., an age belonging to the last two groups) to any user in the dataset. This is also represented in the relatively low mean absolute error of 3.2 years for users who are 40 years old or younger – compared to the mean absolute error of 17.0 years for users above the age of 40.

These results show that for all groups a large share of their users is wrongly classified towards the average age (25.6 years). For the groups [6–17], [31–40], and [41–100] the largest share of their users is assigned an age belonging to the neighboring group in the direction of the average age. Also the remaining three groups – which have the largest share of users assigned to the correct group – have the highest error in the neighboring group towards the average age.

In comparison with existing work (cf. Section 5.6) we additionally showed that predicting whether a user is above or below a certain age can be done with our our algorithm with a reasonably high accuracy. Depending on the actual task such a classification could be used,

Table 5 Misclassifications between age groups

Age	Predicted as					
group	[6–17]	[18–21]	[22–25]	[26–30]	[31–40]	[41–100]
[6–17]	15.8%	**52.5%**	25.8%	4.9%	1.0%	0.0%
[18–21]	5.8%	**41.9%**	41.2%	10.2%	0.9%	0.1%
[22–25]	1.0%	19.9%	**49.7%**	25.9%	3.4%	0.1%
[26–30]	0.0%	4.9%	34.0%	**46.9%**	14.0%	0.1%
[31–40]	0.1%	1.1%	13.1%	**47.7%**	36.8%	1.2%
[41–100]	0.3%	2.0%	6.5%	25.9%	**54.5%**	10.8%

e. g., when recommending music for users in a certain age range. A similar classification could also be used as the basis for a post-processing step, e. g., to detect users that are older than 40 years.

6.3 Gender

During our experiments for gender prediction we created separate subsets containing 50% male and 50% female users. While our algorithm gains a lower accuracy for these follow-up experiments, the results have to be compared to a much lower baseline and show that the algorithm is still able to achieve good results. Table 6 shows the accuracy for male and female users for both of these experiments.

It can be seen that for the experiment on the entire dataset the accuracy for male users (i. e., the majority class) is by far higher than the accuracy for female users (37.6 percentage points difference). For the follow-up experiment the accuracy for male and female users is almost equal (0.5 percentage points difference). These results show that in principle male and female users are detected with a very similar accuracy and that the results depend on the actual dataset and task.

6.4 Country

The results for some of the most common countries are notably better than the overall accuracy for this task – e. g., USA (90.6%), Brazil (90.5%), and Poland (89.9%) – as shown in Fig. 6. However, for Ukraine (31.4%) and Canada (20%) the algorithm reaches rather low accuracy values. Furthermore, within the countries with at least 50 users some countries achieve even low accuracy values – e. g., Austria (0%), Switzerland (0%), and Ireland (3.7%).

When examining the error rates between the countries in the dataset with at least 50 users it can be seen that the algorithm confuses these countries with one specific other country. Table 7 shows all error rates where more than one third of the users from one country with at least 50 users is wrongly classified as belonging to one specific other country. It can be seen that all of these six country pairs share a common border.

Additionally the table includes a similarity rank for each of these country pairs. This rank is calculated by aggregating the similarity values for all country pairs for countries with at least 50 users and sorting them in descending order. The dataset contains 36 countries with 50 or more users, which means there are 630 country pairs and the rank is in the interval [1–630].

Three out of the four most similar country pairs are present in this list, which could explain the high error between these countries. Also the misclassification between Austria and Germany (rank 30) and between Ireland and the United Kingdom (rank 37) could be explained with their high similarity in genre preference. However, the error between Switzerland and Germany cannot be explained with their relatively low similarity rank (rank

Table 6 The prediction accuracy for each gender on the entire dataset and on the balanced datasets

Gender	Prediction accuracy	
	Entire dataset	Balanced datasets
Male	91.7%	77.3%
Female	54.1%	76.8%

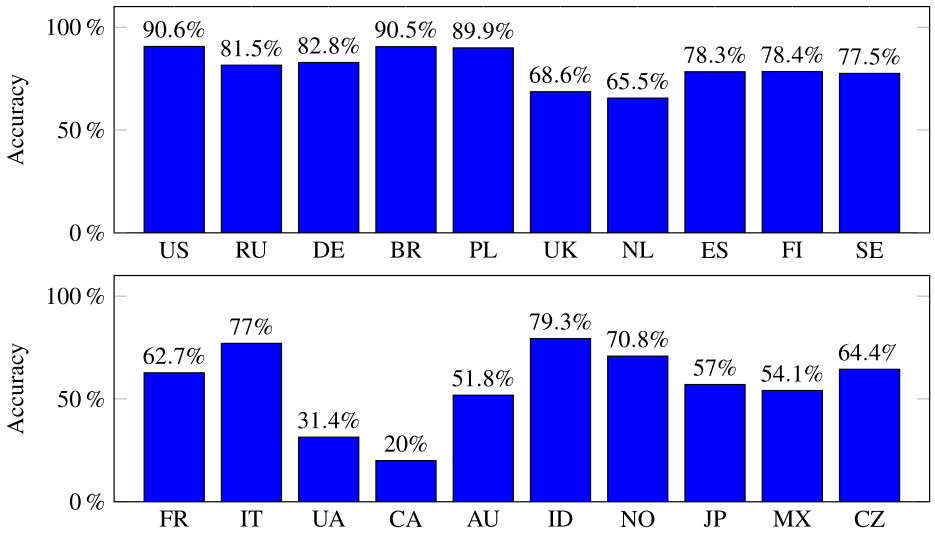

Fig. 6 The prediction accuracy for the 20 most common countries in the dataset

105 out of the 630 country pairs) and seems to be caused by other factors that are not represented by the genre preference similarity. Such factors could be the preference for specific artists or artists with specific attributes (i. e., tags) that is not represented by the genre tags. In fact, besides sharing a border and language with Germany, Switzerland is in itself a diverse country in terms of language and culture, which could result in "grey-sheep" effects in the country profile and a low overall similarity.

In general the error for most countries tends to be towards one specific other country and not spread over multiple countries – one exception is Ireland with high error values towards both the United Kingdom (59.3%) and the United States (24.1%). Out of the 50 most common countries 25 have an error rate above 25% towards one distinct other country and only 10 of these countries have no error rate above 10% towards one distinct other country. For these 50 countries the highest error towards one specific other country affects on average 28.3% of all users from this country.

When allowing errors between the six country pairs in Table 7 (e. g., treating Austrian users that are predicted to be Germans as correct predictions, etc.), the accuracy for this task increases from 69.37% to 73.37%. This accuracy further increases to 75.92%, when we additionally allow errors between the three English-speaking countries USA, United Kingdom, and Australia – which also have highly similar genre preferences.

Table 7 Highest error rates between two countries

Countries	Error	Similarity rank
AT → DE	72.7%	30
BY → RU	70.9%	4
IE → UK	59.3%	37
CA → US	59.0%	2
UA → RU	50.2%	1
CH → DE	46.3%	105

Depending on the application scenario for which the predictions are used, some of these errors are negligible. For example, when using the nationality prediction in the context of recommender systems, it might be possible to generate clusters for these highly similar countries and create recommendations based on these clusters rather than the actual countries (e. g., predict artists that are popular in the German-speaking countries Germany, Austria, and Switzerland).

In summary, finding out a user's nationality (which is unique to our approach when compared to existing work) allows additional applications than when just knowing gender and age. For instance, recommendations can be made area-specific or language-specific, either for a specific country our several countries grouped based on their similarity. Also different legal regulations of different countries could automatically be taken into account when knowing users' nationality (e.g., parental discretion laws in different countries).

7 Use case: improving music recommendation by demographic filtering

To illustrate that demographic information automatically inferred by our approach can actually help improve music recommender systems, we conduct a rating prediction task as follows: we first normalize and scale the playcount values in the user-artist-matrix of the LFM-1b dataset to the range [0, 1000] for each user individually, assuming that higher

Table 8 Root mean square error and weighted mean absolute error employing SVD on the playcounts scaled to [0, 1000], for various demographic user splits

No. Users	User group	RMSE	MAE
120175	All	5.763	2.326
3408	(6,17)	8.322	3.456
13758	(18,21)	6.568	2.383
13194	(**22,25**)	5.521	1.852
7742	(**26,30**)	5.684	1.629
5109	(**31,40**)	3.841	1.366
1661	(**41,50**)	4.098	1.520
595	(**51,60**)	4.043	1.577
301	(**61,100**)	4.056	1.537
10248	**US**	5.236	1.772
5013	**RU**	4.995	1.771
4577	**DE**	4.572	1.694
4533	**UK**	4.343	1.617
4402	PL	8.221	2.623
3877	BR	8.767	3.284
1407	**FI**	4.776	1.717
1375	**NL**	4.053	1.444
1241	**ES**	4.490	1.670
1231	**SE**	3.780	1.264
15781	female	7.174	2.291
39931	**male**	5.114	1.792

For user groups highlighted in bold face font, an improvement can be realized using demographic filtering

numbers of playcounts indicate higher user preference for an artist. We then apply singular value decomposition (SVD) according to [27], equivalent to probabilistic matrix factorization, to factorize the user-artist matrix (UAM) and in turn effect rating prediction. In 5-fold cross-validation experiments, we use root mean square error (RMSE) and mean absolute error (MAE) as performance measures.

To obtain an overall performance score, independent of demographic information, we first conduct an experiment using the set of all users (the full UAM) and report results of the error measures in the first row of Table 8. To investigate the influence of the different demographic characteristics on recommendation performance, we then create respective subsets of users, and perform SVD for each subset individually. Regarding citizenship, we investigate the top 10 countries in the dataset. What we can observe from Table 8 is that such a demographic filtering of users prior to performing rating prediction yields better recommendation performance for the vast majority of user groups.

8 Conclusion and outlook

Our experiments show that the listening history of a person allows to infer certain demographic information (RQ1). All three user traits age, gender, and country can be predicted to a substantial degree. For age the regression algorithm achieves an error that is 33.7% below the baseline error. For the balanced gender prediction and for the prediction of the nationality the increase in accuracy is 54.1% and 264.5% over the baseline, respectively. Even with a very small amount of listening events meaningful predictions can be made (RQ2). With an increasing number of events the performance of the classifiers for all three user trait prediction tasks steadily increases. Finally, we have shown that some of the errors made by the system can be explained, e. g., with the similarity between country pairs (RQ3), and be mitigated, e. g., by grouping these country pairs in recommendations. Some of these factors could be utilized to improve the performance of the classifier.

We have shown that our approach can indeed predict additional information about the users of online music listening services, solely from their listening histories. While the broad categorizations can help in tailoring collaborative as well as content-based recommender systems to their user groups and improve recommendation performance (RQ4), given the shown current limitations, however, it seems unlikely to generally predict personal information about the users that can affect their privacy.

While this extension of our earlier work already tackled some aspects, there are still points left for future work. One area that could be explored are additional features, such as additional information about the artists extracted from different sources or content-based features extracted from the most popular songs, provided the respective audio is available. Additionally, listening session-based analyses could allow to create better features especially for deep networks such as recurrent neural networks.

Acknowledgements This work has been supported by the Christian Doppler Forschungsgesellschaft, Austria, Primetals Technologies, and the Austrian Research Promotion Agency (FFG) under BRIDGE 1 project no. 858514 (*SmarterJam*).

Funding Information Open access funding provided by Johannes Kepler University Linz.

References

1. Cheng Z, Shen J (2014) Just-for-me: an adaptive personalization system for location-aware social music recommendation. In: Proceedings of international conference on multimedia retrieval. ACM, New York, ICMR '14, pp 185:185–185:192. https://doi.org/10.1145/2578726.2578751
2. Cheng Z, Caverlee J, Lee K (2010) You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on information and knowledge management. ACM, pp 759–768. https://doi.org/10.1145/1871437.1871535
3. Cheng Z, Shen J, Mei T (2014) Just-for-me: an adaptive personalization system for location-aware social music recommendation. In: Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. ACM, New York, SIGIR '14, pp 1267–1268. https://doi.org/10.1145/2600428.2611187
4. Cheng Z, Shen J, Hoi SC (2016) On effective personalized music retrieval by exploring online user behaviors. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 125–134
5. Cheng Z, Shen J, Nie L, Chua TS, Kankanhalli M (2017) Exploring user-specific information in music retrieval. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. ACM, New York, SIGIR '17, pp 655–664. https://doi.org/10.1145/3077136.3080772
6. Conover MD, Goncalves B, Ratkiewicz J, Flammini A, Menczer F (2011) Predicting the political alignment of twitter users. In: IEEE Third international conference on privacy, security, risk and trust and IEEE Third international conference on social computing, pp 192–199. https://doi.org/10.1109/PASSAT/SocialCom.2011.34
7. de Montjoye Y, Quoidbach J, Robic F, Pentland A (2013) Predicting personality using novel mobile phone-based metrics. In: Proceedings of the 6th international conference on social computing, behavioral-cultural modeling, and prediction. Springer, pp 48–55. https://doi.org/10.1007/978-3-642-37210-0_6
8. Fuller J, Hubener L, Kim YS, Lee JH (2016) Elucidating user behavior in music services through persona and gender. In: ISMIR, pp 626–632
9. Golbeck J, Robles C, Turner K (2011) Predicting personality with social media. In: Proceedings of the 2011 annual conference on human factors in computing systems. ACM, pp 253–262. https://doi.org/10.1145/1979742.1979614
10. Google (2016) Freebase data dumps. https://developers.google.com/freebase/data
11. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I (2009) The WEKA data mining software: an update. SIGKDD Explorations 11(1):10–18. https://doi.org/10.1145/1656274.1656278
12. Hastie T, Tibshirani R (1998) Classification by pairwise coupling. In: Proceedings of the 1997 conference on advances in neural information processing systems 10. MIT Press. https://doi.org/10.1007/978-3-642-30353-1_3
13. Holmes G, Hall M, Frank E (1999) Generating rule sets from model trees. In: Proceedings of the 12th Australian joint conference on artificial intelligence. Springer, pp 1–12. https://doi.org/10.1007/3-540-46695-9_1
14. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24(6):417–441, 498–520. https://doi.org/10.1037/h0071325
15. Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: Proceedings of the 2008 8th IEEE international conference on data mining. IEEE, pp 263–272. https://doi.org/10.1109/ICDM.2008.22
16. John G, Langley P (1995) Estimating continuous distributions in bayesian classifiers. In: Proceedings of the 11th conference on uncertainty in artificial intelligence. Morgan Kaufmann, pp 338–345
17. Keerthi S, Shevade S, Bhattacharyya C, Murthy K (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput 13(3):637–649. https://doi.org/10.1162/089976601300014493
18. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110(15):5802–5805. https://doi.org/10.1073/pnas.1218772110
19. Krismayer T, Schedl M, Knees P, Rabiser R (2017) Prediction of user demographics from music listening habits. In: International workshop on content-based multimedia indexing. ACM, Florence. https://doi.org/10.1145/3095713.3095722

20. le Cessie S, van Houwelingen J (1992) Ridge estimators in logistic regression. Appl Stat 41(1):191–201. https://doi.org/10.2307/2347628
21. Liu J, Yang Y (2012) Inferring personal traits from music listening history. In: Proceedings of the 2nd international ACM workshop on music information retrieval with user-centered and multimodal strategies. ACM, pp 31–36. https://doi.org/10.1145/2390848.2390856
22. Malmi E, Weber I (2016) You are what apps you use: demographic prediction based on user's apps. In: Proceedings of the Tenth international AAAI conference on Web and social media (ICWSM), pp 635–638
23. Ortigosa A, Carro RM, Quiroga JI (2014) Predicting user personality by mining social interactions in facebook. J Comput Syst Sci 80(1):57–71. https://doi.org/10.1016/j.jcss.2013.03.008
24. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
25. Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C, Smola A (eds) Advances in kernel methods - support vector learning. MIT Press. https://doi.org/10.1023/A:1012474916001
26. Quinlan R (1992) Learning with continuous classes. In: Proceedings of the 5th Australian joint conference on artificial intelligence, World Scientific, pp 343–348
27. Salakhutdinov R, Mnih A (2007) Probabilistic matrix factorization. In: Proc. NIPS, Curran associates Inc., USA, pp 1257–1264
28. Schedl M (2016) The LFM-1b dataset for music retrieval and recommendation. In: Proceedings of the ACM international conference on multimedia retrieval. ACM, pp 103–110. https://doi.org/10.1145/2911996.2912004
29. Schedl M (2017) Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. International Journal of Multimedia Information Retrieval 6(1):71–84. https://doi.org/10.1007/s13735-017-0118-y
30. Schedl M, Flexer A (2012) Putting the user in the center of music information retrieval. In: ISMIR, pp 385–390
31. Schedl M, Flexer A, Urbano J (2013) The neglected user in music information retrieval research. J Intell Inf Syst 41:523–539. https://doi.org/10.1007/s10844-013-0247-6
32. Schedl M, Hauger D, Farrahi K, Tkalčič M (2015) On the influence of user characteristics on music recommendation. In: Proceedings of the 37th European conference on information retrieval. Springer. https://doi.org/10.1007/978-3-319-16354-3_37
33. Shevade S, Keerthi S, Bhattacharyya C, Murthy K (2000) Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Netw (11):1188–1193. https://doi.org/10.1109/72.870050
34. Smola A, Schoelkopf B (1998) A tutorial on support vector regression. Tech. rep., neuroCOLT2 Tech. Rep. NC2-TR-1998-030
35. Su J, Zhang H, Ling C, Matwin S (2008) Discriminative parameter learning for bayesian networks. In: Proceedings of the 25th international conference on machine learning. ACM, pp 1016–1023. https://doi.org/10.1145/1390156.1390284
36. Volkova S, Bachrach Y, Durme BV (2016) Mining user interests to predict perceived psycho-demographic traits on twitter. In: IEEE Second international conference on big data computing service and applications, pp 36–43. https://doi.org/10.1109/BigDataService.2016.28
37. Wang Y, Witten I (1997) Induction of model trees for predicting continuous classes. In: Poster papers of the 9th European conference on machine learning. Springer
38. Wu M, Jang J, Lu C (2014) Gender identification and age estimation of users based on music metadata. In: Proceedings of the 15th international society for music information retrieval conference. ISMIR, pp 555–560
39. Youyou W, Kosinski M, Stillwell D (2015) Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci 112(4):1036–1040. https://doi.org/10.1073/pnas.1418680112
40. Zhuang J, Mei T, Hoi SC, Hua XS, Li S (2011) Modeling social strength in social media community via kernel-based learning. In: Proceedings of the 19th ACM international conference on multimedia. ACM, pp 113–122

Thomas Krismayer is a Ph.D. Student at the Christian Doppler Laboratory for Monitoring and Evolution of Very-Large-Scale Software Systems at Johannes Kepler University Linz, Austria. He holds a Master's degree in Computer Science from the Johannes Kepler University Linz. His research interests include information retrieval, machine learning, data mining, and system monitoring.

Markus Schedl is an Associate Professor at the Johannes Kepler University Linz / Department of Computational Perception. He graduated in Computer Science from the Vienna University of Technology and earned his Ph.D. in Computer Science from the Johannes Kepler University Linz. Markus further studied International Business Administration at the Vienna University of Economics and Business Administration as well as at the Handelshögskolan of the University of Gothenburg, which led to a Master's degree. His main research interests include web and social media mining, information retrieval, multimedia, and music information research.

Peter Knees is Assistant Professor of the Institute of Information Systems Engineering of TU Wien in Vienna, Austria. For over a decade he has been an active member of the Music Information Retrieval research community, reaching out to the related areas of multimedia, text IR, recommender systems, and the digital arts.

Rick Rabiser is a Senior Researcher at the Christian Doppler Laboratory for Monitoring and Evolution of Very-Large-Scale Software Systems at Johannes Kepler University Linz, Austria. He holds a diploma and a Ph.D. in Business Informatics as well as the Habilitation (venia docendi) in Practical Computer Science from Johannes Kepler University Linz. His research interests include variability management, software product lines, software monitoring, requirements engineering, software maintenance and evolution, and usability engineering.