AUTOMATIC MUSIC RECOMMENDATION SYSTEMS: DO DEMOGRAPHIC, PROFILING, AND CONTEXTUAL FEATURES IMPROVE THEIR PERFORMANCE?

ABSTRACT

Traditional automatic music recommendation systems performance typically rely on the accuracy of statistical models learned from past preferences of users on music items. However, additional sources of data such as demographic attributes of listeners, their listening behaviour, and their listening contexts encode information about listeners, and their listening habits, that may be used to improve the accuracy of music recommendation models. In this paper we introduce a large dataset of music listening histories with listeners’ demographic information, and a set of features to characterize aspects of people’s listening behaviour. The longevity of the collected listening histories, covering over two years, allows the retrieval of basic forms of listening context. We use this dataset in the evaluation of accuracy of a music artist recommendation model learned from past preferences of listeners on music items and their interaction with several combinations of people’s demographic, profiling, and contextual features. Our results indicate that using listeners’ self-declared age, country, and gender improve the recommendation accuracy by 8 percent. When a new profiling feature termed exploratoryness was added, the accuracy of the model increased by 12 percent.

1. LISTENING BEHAVIOUR AND CONTEXT

The context in which people listen to music has been the object of study of a growing number of publications, particularly coming from the field of music psychology. Kone˘cni has suggested that the act of music listening has vacated the physical spaces devoted exclusively to music performance and enjoyment long ago, and that music nowadays is listened to in a wide variety of contexts [13]. As music increasingly accompanies our everyday activities, the music and the listener are not the only factors, as the context of listening has emerged as another variable that influences, and is influenced, by the other two factors [11]. It has been also observed that people consciously understand these interactions [6] and use them when choosing music for daily life activities [23]. The context of music listening seems to influence the way in which people chooses music, and so music recommenders should suggest music items to fit the situation and needs of each particular listener. Modelling the user needs was identified by Schedl et al. as one key requirement for developing user-centric music retrieval systems [20]. They noted also that personalized systems customize their recommendations by using additional user information, and context-aware systems use dynamic aspects of the user context to improve the quality of the recommendations. The need for contextual and environmental information was highlighted by Cunningham et al. and others [5, 12, 16]. They hypothesized that listeners’ location, activity, and context were probably correlated with their preferences, and thus should be considered when developing music recommendation systems. As a result, frameworks for abstracting the context of music listening by using raw features such as environmental data have been proposed in the literature [16, 22]. While some researchers have reported that context-aware recommendation systems perform better than traditional ones [15, 22, 24], others have shown only minor improvements [10]. Finally, others have carried out experiments with only the most highly-ranked music items, probably leading to models biased by popularity [15, 25]. We will now discuss the impact of using listeners’ demographic and profiling characteristics— hereafter referred to as user-side features [19]—in improving the accuracy of a music recommendation model. User-side features were extracted from self-declared demographics data and a set of custom-built profiling features characterizing the music listening behaviour of a large amount of users of a digital music service. Their music listening histories were disaggregated into different time spans to evaluate if the accuracy of models changed using different temporal contexts of listening. Finally, models based on latent factors were learned for all listening contexts and all combinations of user-side features. Section 2 presents the dataset collection, Section 3 introduces a set of custom-built features to profile listeners’ listening behaviour, and Section 4 describes the experimental set up and presents the results.

2. DATASET

We are interested in evaluating the impact of using demographic and profiling features, as well as contextual information, for a large number of people, on the prediction accuracy of a music artist recommendation model. A few publicly available datasets for music listening research provide information relating people and music items. Dror et al. presented a dataset of 1M people’s aggregated ratings on music items [7]. McFee et al. introduced a dataset of song playcounts of 1M listeners [17]. Neither of these two datasets, however, provided timestamps of the music logs or demographic information about the listeners. Celma provided a dataset of playcounts with listeners’ demographic data for 360K listeners and a set of listening histories with full time-stamped logs; however this last dataset only included logs for 1K listeners [3]. Cantador et al. presented another small dataset with song playcounts for 2K users [2]. Finally, EMI promised a dataset of 1M interviews about people’s music appreciation, behaviour, and attitudes [9], but only partial information was made available. None of the aforementioned datasets provide, at the same time and for a large amount of listeners, access to full music listening histories as well as people’s demographic data. This means it is not possible to extract all of the userside features that we were interested in, and so we decided to collect our own dataset made with music listening histories from Last.fm. Last.fm stands out from most online digital music services because it not only records music logs of songs played back within its own ecosystem, but also from more than 600 media players. Next, we will present the criteria and acquisition methods used to collect a large number of music listening histories from the Last.fm service.

2.1 Data criteria, acquisition, and cleaning

Aggregating people’s music listening histories requires collapsing their music logs into periods of time. In order to obtain data evenly across aggregated weeks, months, seasons, or years, we searched for listeners with an arbitrary number of at least two years of activity submitting music logs since they started using the system, and also with an average of ten music logs per day. These two restrictions forced our data gathering crawler to search for listeners with a minimum of 7,300 music logs submitted to the Last.fm database. Also, these constraints assured us that we would collect listening histories from active listeners with enough data to perform a good aggregation over time. Data acquisition was performed by means of using several machines calling the Last.fm API during a period of two years (2012–2014). We collected listening histories by using the Last.fm’s API method user.getRecentTracks(). This API call allowed us to obtain full listening histories. Along with this data, we also stored all available metadata for each listener, including the optional self-declared demographic features: age, country, and gender. We performed several processes of data filtering and cleaning in order to avoid noisy data. For example, we realized that there were listeners with numerous duplicated music logs (i.e., same timestamp for many music item IDs), and listening histories with a great deal of music logs that were too close in time (i.e., less than 30 seconds apart, which is the minimum that Last.fm requires to consider a played track as a valid music log). Hence, we decided to filter out all duplicated logs as well as logs that were less than 30 seconds apart in time.

2.2 Dataset demographics

Our dataset consists of 27 billion music logs taken from 594K users’ music listening histories. This large repository of music listening records accounts for the interaction of listeners with more than 555K different artists, 900K albums, and 7 million tracks. There are music listening histories from people in 239 self declared different countries, with listeners from all time zones represented. However, listeners from Africa, South Asia, and East Asia are under-represented in our dataset. In fact, the 19 “top countries” combined account for more than 85 percent of the total number of listeners in the dataset. Table 1 summarizes some of the overall and demographic characteristics of users in the dataset.

Table 1. Dataset summary (Demographic: the percentage of people who provided demographic information)

Table 1 shows that large proportion of listeners selfdeclared their age, gender, and country. Previous research on online profiles concluded that people usually want to be well typified by their online profiles [4], and so we assumed there is a high degree of truth in these demographic features. Listeners from all ages are not equally represented in the dataset. The age distribution is biased towards young people, with an average age of 25 years old.

3. FEATURES FOR LISTENER PROFILING

We hypothesized that by better understanding the listening behaviour of people, we will be able to more accurately model the user needs. Hence, the recommendation can be tailored to each listener and the prediction accuracy will probably improve. A set of computational features that attempt to describe some aspects of music listening behaviour in relation to musical artists was already proposed in previous research

[21]. However, the ranking of the music items was not take into consideration and feature values were binned into categories. In our approach we try to represent similar characteristics of listening behaviour but we also consider the position of the music items within each listener’s ranking as well as using normalized feature values to express the precise value of a certain listening behaviour characteristic in relation to a music item.

3.1 Feature design

We restricted ourselves to designing three novel features to describe listener behaviours: exploratoryness, mainstreamness, and genderness. Values for these features were computed for the three types of music items in the dataset: tracks, albums, and artists. Therefore, each listener’s listening behaviour was described by a vector of nine values. We will describe the goals behind each one of these features, give details about their implementation, visualize data patterns, and provide some analysis about the results.

3.1.1 Exploratoryness

To represent how much a listener explores different music instead of listening to the same music repeatedly we developed the exploratoryness feature. For each user x’s listening history, let L be the number of submitted music logs, Sk be all submitted music items of type k, where k={tracks, albums, artists}, sk,i be the number of music logs for the given music item key k at ranking i. We computed the exploratoryness ex,k for listener x on a given music item of type k as: (CALCUL)

Exploratoryness returns a normalized value, with values closer to 0 for users listening to the same music item again and again, and values closer to 1 for users with more exploratory listening behaviour.

3.1.2 Mainstreamness

With the goal of expressing how similar a listener’s listening history is to what everyone else listened to, we developed the mainstreamness feature. It analyses a listener’s ranking of music items, and compares it with the overall ranking of artists, albums, or tracks, looking for the position of co-occurrences. For each user x’s listening history, let N be the number of logs of the music item ranked first in the overall ranking, L be the number of submitted music logs, Sk be all submitted music items of type k, where k={tracks, albums, artists}, sk,i be the number of music logs for the given music item key k at ranking i, and ok,i be the number of music logs in the overall ranking of music item type k ranked at position i. We defined the mainstreamness feature mx,k for listener x on a given music item of type k as: (CALCUL)

Listening histories of people with a music item’s ranking similar to the overall ranking receive mainstreamness values closer to 1. Listeners’ mainstreamness whose ranking differ more from the overall ranking receive values closer to 0.

3.1.3 Genderness

With the aim of expressing how close a listener’s listening history is to what females or males are listening to, we developed the genderness feature. The genderness feature computation basically relies on mainstreamness, but instead of computing just one overall ranking from all listeners, it uses two rankings: one made with music logs from listeners self-declared as female, and another one from male data. For each user x’s listening history and music item of type k, let mx,k,male be the mainstreamness computed with the male ranking, mx,k,female be the mainstreamness calculated with the female ranking. We defined the feature genderness gx,k for listener x on a given music item of type k as: (CALCUL)

3.2 Profiling listeners

To illustrate how the features we developed can be used to profile listeners, we calculated exploratoryness, mainstreamness, and genderness of users in our dataset. In order to not violate the homogeneity of variance we binned listeners into four age groups with balanced number of samples for each group. To obtain balanced groups, we drew a random sample of 100 people of each age, and created 10-year groups with 1000 people each. We then bootstrapped these groups with 1000 replications of the original sample and calculated 95 percent CI error bars. Although we quantified these characteristics in the relation of listeners with artists, albums, and tracks, and their interaction with listeners’ age group, preliminary tests indicated that the interaction with artists was most significant. Therefore, for the rest of the paper we present only the results of the interaction between listeners and artists. Figure 1 shows feature means by age group as well as 95 percent CI bars. In terms of artist exploratoryness, Figure 1(a) shows that while younger listeners in our dataset tend to listen more often to the same performers than adults, older listeners tend to explore more artists. Also, the rise in exploratoryness tends to stabilize in the midthirties. Figure 1(b) shows that while younger people listen more to the same artists that everyone is listening to, older people tend to listen to less common performers. This effect could be generated by the behaviour of older people or the fact that there are fewer older people in the original dataset, and so the artists they listen to are less represented in the overall ranking. Figure 1(c) shows artist genderness by age and gender. Listeners self-declared as male tend to listen more to music that is ranked higher in the male ranking, in all age groups, however their preference for the male ranking diminishes with age. Females, on the contrary, listen more to artists ranked higher in the female ranking when they are young, but adult women listen more to artists ranked higher in the male ranking. Overall, men and women have opposite trends of genderness in the different age groups, which seem to stabilize as they mature.

Figure 1. Feature means and 95% CI bars for a random group of listeners in our dataset. Each age group has 1K listeners. Error bars were calculated by taking 1K populations replicated from the original sample using bootstrap. (a) Artist exploratoryness by age group of listeners, (b) artist mainstreamness by age group of listeners, and (c) artist genderness by listeners’ age group and gender.

We hoped that the aforementioned features captured some information about people’s listening behaviour and will help to improve the accuracy of a music recommender model. However, as genderness was derived directly from mainstreamness, we did not use it in the experimental procedure for evaluating a music recommendation model with user-side data.

4. EXPERIMENTAL PROCEDURE

Our goal is to evaluate if demographics, behavioural profiles, and the use of observations from different contexts improve the accuracy of a recommendation model. Our sources of data involve a matrix of user preferences on artists derived from implicit feedback, a set of three categorical demographic features for each user: age, country, and gender, and a set of two continuous-valued features for describing people’s listening behaviour: exploratoryness and mainstreamness. Preference matrixes were generated by considering full week of music listening histories data, as well as data coming from music logs submitted on weekdays and weekends only. We followed a similar approach to Koren et al., in which a matrix of implicit feedback values expressing preferences of users on items is modelled by finding two lower dimensional matrixes of rank f Xn×f and Ym×f , which product approximates the original preference [14]. The goal of this approach is to find the set of values in X and Y that minimize the RMSE error between the original and the reconstructed matrixes. However, this conventional approach of matrix factorization for evaluating the accuracy of recommendation models using latent factors does not allow the researcher to incorporate additional features, such as user-side features. In order to incorporate latent factors as well as user-side features into one single recommendation model, we used the Factorization Machines method for matrix factorization and singular value decomposition [18]. In this approach, interactions between all latent factors as well as additional features are computed within a single framework, with a computational complexity that is linear to the number of extra features. In order to perform a series of experiments with different sets of model parameters and user-side features in a timely fashion, we randomly sampled 10 percent of peruser music listening histories in the dataset, and we split this new subset into two disjoint sets: training (90 percent) and testing (10 percent) datasets. The training dataset had more than 60M observations from 59K users on 432K artists, with a density of observations of about 0.24 percent.

We aggregated each dataset of listening histories by creating <user, artist, playcounts> triples. Then,

we transformed the number of playcounts in each triple into a 1–5 Likert scale value by means of calculating the complementary cumulative distribution of artists per listener [3]. Hence, artists in each distribution quintile were assigned with a preference value according to how much each user listened to them. In order to learn the best set of parameters of the recommendation model, we performed a grid search on the \_ regularization parameter as well as the f number of latent factors with no user-side data, just using plain matrix factorization for the matrix of preferences of users on artists. Finding a good \_ value helps to avoid overfitting the observed data by penalizing the magnitudes of the learned parameters. Finding the best f number of factors helps to obtain a better recommendation accuracy while also providing a set of to-be-interpreted latent factors. We used the Graphlab Create framework 1 to search over the number of latent factors in the range [50, 200] , and regularization values in the range [1×10-5, 1×10-8]. The best combination of parameters was achieved for \_=1×10-7 and f=50 latent factors. We used the Adaptive Stochastic Gradient Descent optimization algorithm [8] and set the maximum number of iterations at 50.