30Music listening and playlists dataset

Roberto Turrin
ContentWise R&D
roberto.turrin@contentwise.tv

Massimo Quadrana
DEIB, Politecnico di Milano
massimo.quadrana@polimi.it

Andrea Condorelli
ContentWise R&D
andrea.condorelli@contentwise.tv

Roberto Pagano
DEIB, Politecnico di Milano
roberto.pagano@polimi.it

Paolo Cremonesi
DEIB, Politecnico di Milano
paolo.cremonesi@polimi.it

ABSTRACT

We introduce the *30Music* dataset[1], a collection of listening and playlists data retrieved from Internet radio stations through Last.fm API. In this paper we describe the creation process, its content, and its possible uses. Attractive features of the *30Music* dataset that differentiate it from existing public datasets include, among the others, (i) the user listening sessions complete of contextual time information, (ii) the user playlists, and (iii) the positive user ratings, key information to experiment with the task of modeling user taste and recommending playlists.

1. INTRODUCTION

Several challenges in the music domain have been only partially explored due to the scarcity of data available to researchers for experiments. For instance, tasks such as user modeling and playlist recommendation require implicit contextual information about listening events (e.g., user, track, time, duration), explicit information about user preferences (e.g., loved tracks, playlists), and user listening sessions.

In this paper we introduce the *30Music* dataset, a freely-available music dataset designed to overcome these problems. The main innovative aspects of the *30Music* dataset with respect to the existing public datasets are:

- the dataset contains both implicit play events and explicit user ratings (i.e., preferred tracks);
- the dataset contains user-generated playlists;
- play events are organized into listening *sessions*;
- whenever a user plays a track from a playlist, the play event is tagged.

The rest of the paper is organized as follows. Section 2 discusses the existing music datasets. Section 3 presents the process implemented to crawl the data from Last.fm and create the dataset, whose main characteristics are explored in Section 4. Finally, Section 5 draws the conclusions and discusses future work.

2. RELATED DATASETS

There exist a number of publicly-available music datasets, used in several music experiments. Most datasets provide content information (e.g., metadata, tags, acoustic features), but only a few report some user-system interactions (e.g., ratings, play events) useful to profile users and to experiment with personalization tasks.

The Million song dataset (MSD) [1] is a public collection well-known for its size. In fact, it contains audio features (e.g., pitches, timbre, loudness, as provided by the Echo Nest Analyze API[2]) and textual metadata (e.g., Musicbrainz[3] tags, Echo Nest tags, Last.fm tags) about 1M songs (related to 44K artists).

Celma [2] has published two music datasets collected from Last.fm API: 1K-user and 360K-user. The smallest one - 1K-user dataset - contains the user listening habits (20M play events) of less than 1K users. On the other hand, the biggest one - the 360K-user dataset - collects the information about 360K users, but it does not have any listening data other than the number of times a user has listened to an artist. Data are provided as downloaded from the Last.fm API.

Yahoo! Labs have released several music datasets[4]. For instance, the R1 and the R2 datasets provides ratings on artists and songs, respectively, but not user play events.

Some datasets have been extracted from microblogs, such as the Million Musical Tweets Dataset [3]. Finally, The Art of the Mix Playlist dataset[5], consists of 29K user-contributed playlists, containing 218K distinct songs for 60K distinct artists. However, there are no user listening events.

3. DATASET CREATION

The *30Music* dataset has been obtained via Last.fm public API[6]. Last.fm provides free API to track details of user listening sessions. In the case a user has connected his supported player to his Last.fm account, the player "*scrobbles*" the user listening activity, i.e., it transfers the play event to Last.fm that records such user interaction. It is worth noting that only listening events are recorded, while pause/skip events are not scrobbled from the user player to Last.fm, as well as any playlist or explicit preference defined or expressed in the player. The main way for a user to create a playlist in Last.fm is to access to the website; similarly, the user can express explicit preferences ('love') about tracks directly in the website. As a consequence, explicit ratings and playlists stored in Last.fm are not biased by external systems (e.g., the recommendations proposed in the player).

[1] http://recsys.deib.polimi.it/?page_id=54

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

[3] https://musicbrainz.org/

[4] http://webscope.sandbox.yahoo.com/catalog.php

[5] http://labrosa.ee.columbia.edu/projects/musicsim/aotm.html

[6] http://www.last.fm/api

RecSys 2015 Poster Proceedings, September 16-20, 2015, Austria, Vienna.

unfortunately, last.fm requires to scrobble only tracks played for at least half their duration (or for at least 4 minutes), so events not matching these conditions - such as skip events - are not confident (although many events less than 5 seconds have been found in the collected data).

To build the *30Music* dataset, we started from a list of 2M Last.fm usernames from the Chris Meller dataset [7]. Given the list of users, we retrieved their playlists (`User.getPlaylists`) together with the single tracks composing the playlists (`Playlist.fetch`). Starting from users with at least one playlist (about 90K users), we retrieved (`User.getRecentTracks`) the user listening events over a 1-year time window (from Mon, 20 Jan 2014 09:24:19). The raw playlists and user listening events have been enriched with additional information both about users (`User.getInfo`) and tracks (`Track.getInfo`).

Furthermore, the data downloaded with the Last.fm API has been processed using Python scripts exploiting some Apache Spark functions for a distributed processing of the massive amount of data. In order to keep only complete and reliable data, we discarded users with some missing data (e.g., if the track scrobbled by the user has the wrong metadata and it is not recognized by Last.fm, the whole user is discarded). In this way, we maintained only the half of the users that have complete information.

Finally, we defined a new entity, the *user play session*, as an ordered list of play events that are assumed to be consequently listened by the user with no interruptions. We define a play event to be part of a session if it occurs no later than 800 seconds after the previous user play event. This processing required, for each user listening event, to compute the play time, together with the ratio of track effectively listened by the user.

30Music format.

The *30Music* dataset is released in accordance with the [anonimized for double blind review] data format, a multigraph representation oriented to recommender system evaluation that explicitly represents *entities* (i.e., nodes) and *relations* (i.e., edges).

Entity model any object that can be recommended. The dataset is formed by 45K users, 5.6M tracks, 50K playlists, 600K artists, 200K albums, and 280K tags. Relations model links between two (or more) entities. We defined 31M user play events, 2.7M user play sessions, and 4.1M user love preferences.

4. DATASET ANALYSIS

The dataset contains 31,351,954 play events organized into 2,764,474 sessions (an average of 11 play events per session). The dataset contains also 4,106,341 explicit ratings (loved tracks), with an average of 33 ratings per user, and 57,561 user-created playlists. The number of events without track duration is 1,277,893 (4.08%).

We can observe that play events present a moderate long-tail distribution: the 20% most popular tracks collect 80% of the play events. This long tail effect is mitigated by focusing on preferred tracks (i.e., loved tracks and tracks in the playlists). We observe that the same percentage of play events (80%) involves twice the tracks (40%) when considering tracks in the playlists. We can deduce that users have preferences spanning many different tracks, but their listening behaviour is biased toward the most popular tracks.

A similar analysis has been performed by aggregating the tracks of the same artist. Differently from tracks, these play events present a strong long-tail distribution: the 20% most popular artists collect more than 95% of the play events. This long tail effect is strongly mitigated when analyzing preferred tracks. The same percentage of play events (95%) involve 50% of the artists when considering tracks in the playlists. We can deduce that users have preferences spanning many different artists, but their listening behaviour is strongly biased toward the most popular artist.

An analysis of the empirical cumulative explicit like distribution as a function of the (percentage) number of tracks and artists shows that only the 14.73% of the tracks and the 19.93% of the artists have received at least one explicit preference. We observed that the distribution of the explicit ratings within tracks does not exhibit a strong long-tail behavior. The 5% of the most popular tracks collect the 75% of the explicit ratings. On the other hand, the number of explicit ratings is strongly skewed toward a few very popular artists. The 5% of the of most popular artists collect more than the 90% of the explicit ratings. These results confirm our previous intuitions over users' listening behaviour. Users tend to love (and to listen to) a few very popular artists. However, their preference spans across several tracks of these very popular artists. Still, they tend to provide explicit rating for few of the tracks they have listened to. This can be due to the mechanism adopted by Last.fm to collect explicit feedback, which forces users to move from their usual music player, to access to the Last.fm online service and to provide their "love" to a track there. This clearly imposes a heavy burden over users, but on the other hand it enhances the strength of each explicit rating, because it is a clear expression of the willingness of the specific user to provide that feedback.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the *30Music* dataset, a music dataset consisting of both user interactions (i.e., user play sessions) and user explicit preferences (i.e., playlists, preferred tracks). The dataset is made available to the research community and we expect it will foster the exploration of the several challenges still open in the settings of online music applications.

Acknoledgements.

The research leading to these results was performed in the CrowdRec project, which has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement n. 610594.

6. REFERENCES

[1] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. *12th Int. Conf. on Music Information Retrieval (ISMIR)*, 2011.

[2] O. Celma. *Music Recommendation and Discovery in the Long Tail.* Springer, 2010.

[3] D. Hauger, M. Schedl, A. Kosir, and M. Tkalcic. The million musical tweet dataset - what we can learn from microblogs. In *Proc. of the 14th Int. Society for Music Information Retrieval Conference*, Nov 4-8 2013.

[7] https://opendata.socrata.com/Business/Two-Million-LastFM-User-Profiles/5vvd-truf