Musical Track Popularity Mining Dataset

Ioannis Karydis[1], Aggelos Gkiokas[2], and Vassilis Katsouros[2]

Dept. of Informatics, Ionian University, Kerkyra 49132, Greece[1]
karydis@ionio.gr
Institute for Language & Speech Processing, Athena RIC, Athens 15125, Greece[2]
{agkiokas,vsk}@ilsp.gr

Abstract. Music Information Research requires access to real musical content in order to test efficiency and effectiveness of its methods as well as to compare developed methodologies on common data. Existing datasets do not address the research direction of musical track popularity that has recently received considerate attention. Existing sources of musical popularity do not provide easily manageable data and no standardised dataset exists. Accordingly, in this paper we present the Track Popularity Dataset (TPD) that provides different sources of popularity definition ranging from 2004 to 2014, a mapping between different track/ author/ album identification spaces that allows use of all different sources, information on the remaining, non popular, tracks of an album with a popular track, contextual similarity between tracks and ready for MIR use extracted features for both popular and non-popular audio tracks.

1 Introduction

One of the most important requirements of Music Information Research (MIR) is access to pertinent musical content. The experimentation on this content mostly aims on the testing of the efficiency and effectiveness of the MIR methods, while providing reference for comparison of new and existing methods in order to show progress. In rare cases, the use of synthetic data can be helpful to the aforementioned use of data in MIR experiments, though music, being highly an artistic form of expression, does not always adhere to a set of deterministic rules that researchers could rely on in order to avoid the requirement for access to real musical content.

Accordingly, in MIR, as in most areas of scientific research, the collection, distribution and use of datasets is of great importance, despite the litany of legal issues [6] that may arise from such practices. Music data for the purposes of MIR usually refer to audio files of recorded performed musical pieces, symbolic representation of a piece, lyrics, metadata as well as contextual to the piece information mainly collected through social networks pertaining to the users' perception of or activities on the pieces. Thus, following the need for such content exchange and its intended use, MIR datasets additionally include commonly used derivative transformations of all the aforementioned musical information in order

to avoid legal implications as well as to spare users of time and resources required for these to be produced.

Numerous datasets exist in MIR [9] that cover a broad area of the domain, though none is immediately applicable for knowledge extraction from the popularity that musical pieces receive. The process of track popularity prediction prior to or during the initial period of a track's release has long been a requirement of the musical industry. Interestingly enough, the gains of such a prediction go far beyond the obvious benefits of allowing musical labels to identify financially interesting clients, as the whole ecosystem (artists and listeners) also profits. Despite the aforementioned benefits, it was only after the commercial application by Polyphonic HMI [1] that the issue gained significant attention as a research direction, as early as 2005 [4].

1.1 Motivation and Contribution

Existing commonly used services, such as Spotify[2], Billboard[3] and Last.fm[4] that provide popularity of musical content do not offer easily manageable data. Spotify's localised charts, although provided an Application Programming Interface (API), have temporarily according to the service's community helpdesk, ceased to function as of approximately March 2015 and are still offline. Last.fm's localised charts do not offer an API, though Last.fm does indeed provide the aggregated number of listeners and playcounts for all available tracks. Billboard's Hot 100 Chart does not offer an API but does provide the most long termed archives, dating back to August 9th, 1958.

To add to the difficulties of collecting track popularity information, the aforementioned services utilise their respective track identification space making collective use of multiple popularity sources rather difficult. Moreover, collecting just the tracks that exceed the popularity threshold, research cannot deal integrally with the separation of hits from non-hits as no information on non-hits is available, since the collected information only contains the degree of popularity. Finally, having access to the content of the audio files of the popularity chart is, among other parameters, very important in the selection of the track's representative features that will lead to high quality predictions.

To address these requirements, we introduce the Track Popularity Dataset (TPD), a collection of track popularity data for the purposes of MIR, containing:

1. different sources of popularity definition ranging from 2004 to 2014,
2. information on the remaining, non popular, tracks of an album with a popular track,
3. a mapping between different track/author/album identification spaces that allows use of all different sources,
4. contextual similarity information between all popular tracks,

[1] `http://polyphonichmi.blogspot.gr/p/about-company.html`
[2] `https://charts.spotify.com`
[3] `http://www.billboard.com/charts/hot-100`
[4] `http://www.last.fm/charts`

5. ready for MIR use extracted features for popular & non-popular audio tracks,

The rest of the paper is organised as follows: Section 2 presents background information on Hit Song Science and related work, while Section 3 discuses the proposed dataset, its creation processes as well as a detailed analysis of its content. Next, Section 4 details future directions concerning the dataset that could ameliorate is usability and further support MIR research. Finally the paper is concluded in Section 5.

2 Background and Related research

2.1 Hit Song Science

Hit Song Science (HSS) [13] refers to the MIR direction aiming in predicting the popularity of musical tracks, as presented in top-charts. A number of scenarios' parameters exist as to the prediction's prerequisites, such as the little or no availability of early popularity information, the granularity of popularity definition, the type of input sources representing the musical tracks and many others.

Similarly, under the auspices of HSS numerous research tasks also take place: popularity pattern modelling, binary (hit/non-hit) or otherwise granulated popularity classification, tracks' future position on the popularity chart prediction given current position, popularity correlation to other activities (i.e. twitter posts, music search/download in peer-to-peer networks, etc), prediction of the popular track subset of an album and many more.

The ability to predict the popularity of musical tracks is of great importance to all parties involved in the musical content lifecycle. Creators can work reversely the process of HSS and focus on characteristics that make their songs more probable to be popular in addition to customised characteristics of listeners, markets or distribution channels. The music industry, aiming at maximum profit, could benefit by selecting the most promising of the works for publication as well as, given that popularity predictions can be attributed to specific profile candidate consumers, modify accordingly its marketing plans. Finally, music consumers indirectly increase enjoyment of listening by receiving music that the distribution channels have either selected to fit their profile or that is in general more probable to be of high popularity and thus more probable widely liked.

It is widely claimed that the breadth of characteristics that lead to the popularity of a musical piece exceeds the *per se* track's content i.e., the audio and lyrics. Factors such as artist preferential attachment [1], society and culture [4], the changing musical tastes leading to evolving popularity pattern [11], psychological parameters on the reasons for preferring a track and listening exposure to tracks [12], the video clip of the track [3] just to name a few, play also an important role.

Nevertheless, existing research in the area agree that, beyond the very hard to measure characteristics, quantifiable qualities of musical tracks that contribute

to a track's popularity do exist. Accordingly, the burden remains with the transformation representations of musical tracks that need to adhere to the track's popularity pertinent attributes.

2.2 Existing Research

In the first work on the area, Dhanaraj and Logan [4], utilise SVM and boosting classifiers on both acoustic and lyric information for the purposes of hit songs' separation from non-hits. Their aim is to determine if such a task is feasible or if hit song science claims are to be deemed as impossible, arriving after experimentation at the former.

Chon et al. [3] research for meaningful patterns within musical data while also attempting to predict both how long an album will stay in chart as well as a new album's position in chart on a certain week in the future using with the first few weeks' sales data. The results presented therein indicate interesting correlations.

Pachet & Roy, in [13], and Pachet, in [12], describe a large scale experiment aiming at the validation of current state-of-the-art methods' capabilities to predict the popularity of musical titles based on acoustic and/or contextual features. Both these works suggest that the commonly used features for music analysis are not informative enough to offer judgment on notions related to subjective aesthetics.

In [1], Bischoff et al. propose the music pieces' success prediction by exploitation of social interactions and annotations leaving out content characteristics of the musical tracks. Thus, their method relies on data mined from the Last.fm [5] music social network and the relationship between tracks, artists and albums while reaching promisingly improved results.

In a differentiated scenario, the work of Koenigstein et al. [8] compares peer-to-peer file sharing information on songs to their popularity, as described on the Billboard [6] charts. Their work indicates popularity trends of songs on the Billboard having a strong correlation to their respective popularity on peer-to-peer network. Based on this result, Koenigstein et al. propose a methodology that utilises the aforementioned correlation in order to to predict a songs' success on the popularity charts offering a 2-3 weeks prior to chart announcement prediction with high accuracy.

Following the work [13], Ni et al. [11] on a slightly alternated research question argue the feasibility of popularity prediction, "given a relevant feature set". Based on that work, the website "Score a hit" [7] was also created.

The work of Kim et al. [7] proposes the collection of users' music listening behaviour from Twitter, based on music-related hashtags, for the purposes of predicting popularity rankings. The results reported show high correlation between users' music listening behaviour on Twitter and general music popularity trend.

[5] `http://www.last.fm/`
[6] `http://www.billboard.com/`
[7] `http://www.scoreahit.com`

Singhi & Brown [14] propose a hit detection model based Bayesian networks on solely lyrics' features.

Finally, Burgoyne et al. [2] present a close to the theme of musical track popularity work studying musical content's "catchiness", or the "long term musical salience" of a piece. Despite the broader scope of the musical popularity prediction task, the correlation of catchiness to popularity is evident although most probably one directional, since numerous less-memorable top-chart tracks do exist.

	music representation							Dataset size	Hit definition	Top charts time span
Research	content-based audio	lyrics	subjective contextual	objective contextual	objective metadata	P2P queries	album sales data			
[4]	✓	✓						1700 songs	Billboard top 1	Jan 1956 - Apr 2004
[3]							✓	291 albums	Billboard top 1-25	Sep 2002 - Jun 2006
[12], [13]	✓		✓		✓			32000 songs	HiFind popularity label: low, medium, high	?
[1]			✓					50555 songs	Billboard top 1, 3, 5, 10, 20, 30, 40, 50	Aug 1958 - Apr 2008
[8]						✓		185598176 p2p queries, 200 songs	Billboard top 10, 20, 20, 30, 40, 50, 100	Jan 2007 - Jul 2007
[11]	✓							5000 songs	Billboard top 5	1962-2011
[14]		✓						6815 songs	Billboard top 15, 25, 35	2008-2013
[7]				✓	✓			1806438 tweets, 168 songs	Billboard top 10, 20, 30, 40, 50	Nov 2013 - Jan 2014

Table 1. Existing HSS research dataset details.

The aforementioned existing research, with the exception of [13] and [12], have utilised different datasets to perform experimentation. The diversity of the utilised datasets in terms of size vary greatly as shown in Table 1.

3 The Dataset

The TPD is a collection of information revolving around the notion of track popularity. Its aim is to provide an easy to use collection of information for the purposes of track popularity data mining research tasks. In this Section we detail the creation process and content of the TPD.

3.1 Creation Process

In order to create the TPD, we separated the potential information sources into three distinct categories: the popularity sources, the metadata/content sources and the contextual similarity source.

The selection of the popularity periods was made based on the availability of both popularity information from the sources and access to the tracks' content. Thus, from popularity sources Last.fm and Spotify we collected all available popularity charts at the time of collection, that is from 17 September 2006 up to 28 December 2014 and 28 April 2013 up to 18 January 2015, respectively. From popularity source Billboard we collected the last 10 years, ranging from 03 January 2004 up to 24 January 2015.

Following the collection of the popular tracks from the popularity sources, we utilised the metadata/content sources Apple[8], Spotify[9], 7digital[10] in order to identify and get information for albums of the collected popular tracks and then to gather information on remaining, non-popular, tracks of each album.

Access to the content of the collection's tracks was based on the metadata/content sources' (Apple, Spotify and 7digital) 30 second previews clips as all three web services provide an API for the purposes of searching and streaming the audio clips. The collected files were converted to an appropriate format in order to undergo feature extraction.

While performing the above mentioned information collection processes, it was confirmed that multiple identification spaces do indeed exist for all track/ author/ album entities. Accordingly, and in order to facilitate the interoperability of the collected information, we performed exact match searches in all sources producing thus a mapping between different track/ author/ album identification spaces. As not all sources engulf information on all collected data, the mapping is not complete, but nevertheless, far from sparse (∼55% of the matrix cells contain values). Content for the mapping was collected from both popularity sources and metadata/content sources.

To enrich further the TPD, we additionally included contextual information as to the similarity of the collection's tracks based on Last.fm's API *track.getSimilar* method that provides similarity between tracks, based on listening data.

Finally, for each track of the TPD, three feature-sets extracted directly from the audio content are included in matlab variable MAT-files. The first feature-set, *feature-set A*, is based on jAudio [10] and contains only single overall average and standard deviation values performed on all values of the features over all windows with window size 512 samples and 0% overlap between successive windows. The second feature-set, *feature-set B*, was created with MIRToolbox offering per window feature extraction with window size 1024 samples and 50% overlap between successive windows. The two feature-sets provide different levels of detail on the audio content in order to suit a broad range of applications. The third feature-set, *feature-set C* is based on the periodicity function of the tempo estimation method presented in [5].

3.2 The Content

The TPD contains 23.385 tracks of which, 9.193 are designated as popular by appearing in any of the popularity sources charts, while 14.192 are tracks that appear in one of the 1.843 albums of the popular tracks and are not designated as popular by any of the popularity sources. The popularity ratings records, contain the position of a track for a specific week, collected from Billboard are 57.800, while for Last.fm and Spotify are 43.300 and 6.500, respectively. Of the

[8] `https://www.apple.com/itunes/affiliates/resources/documentation/itunes-store-web-service-search-api.html`

[9] `https://developer.spotify.com`

[10] `http://developer.7digital.com/`

popular tracks, 1,5% are designated in all three sources of popularity, 5,9% in two sources and 92,6% in just one source. The discrepancy in proportions is due to the range of available data by the popularity sources. As far as the contextual similarity based on Last.fm's API *track.getSimilar* method is concerned, 78% of the popular tracks of the dataset have a degree of contextual similarity to other popular tracks of the dataset. As not all tracks' audio files were possible to be found, the TPD contains audio derived features for ∼74% of the tracks.

Of the three feature-sets included in the TPD described in Section 3.1, *feature-set A* is meant as a small, less detailed feature set for fast and simple research applications. The features included in *feature-set A* are: overall standard deviation & overall average of spectral centroid (dimension: 1), spectral rolloff point (dim: 1), spectral flux (dim: 1), compactness (dim: 1), spectral variability (dim: 1), root mean square (dim: 1), fraction of low energy windows (dim: 1), zero crossings (dim: 1), strongest beat (dim: 1), beat sum (dim: 1), strength of strongest beat (dim: 1), strongest frequency via zero crossings (dim: 1), strongest frequency via spectral centroid (dim: 1), strongest frequency via FFT maximum (dim: 1), MFCCs (dim: 13), LPCs(dim: 10), method of moments (dim: 5), partial based spectral centroid (dim: 1), partial based spectral flux (dim: 1), peak based spectral smoothness (dim: 1), relative difference function (dim: 1), area method of moments (dim: 10). The second feature-set, *feature-set B*, contains windowed MFCCs (dim: 13), rolloff (dim: 1), brightness (dim: 1), flux (dim: 1), zero crossings (dim: 1), inharmonicity (dim: 1), centroid (dim: 1), spread (dim: 1), skewness (dim: 1), kurtosis (dim: 1), flatness (dim: 1), entropy (dim: 1). The third feature-set, *feature-set C*, contains 276 target tempi. For each target tempo this feature-set contains eight energy bands and one chroma (dim: 9).

In order to provide an aggregated glimpse of the popularity records of the dataset by contrasting the popularity sources, Figure 1 shows the normalised probability density of next week's rank increase/decrease (position change) given current position for all three sources of popularity. Moreover, Figures 2 and 3 show the probability density of rank position when entering and leaving respectively the top-100 popularity chart for all three sources of popularity.

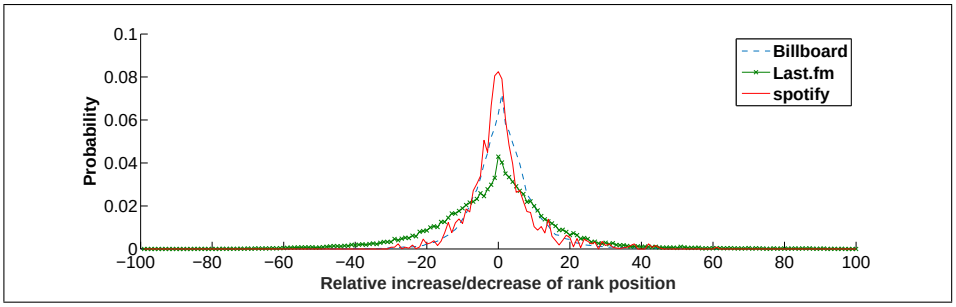

Fig. 1. The normalised probability density of position change.

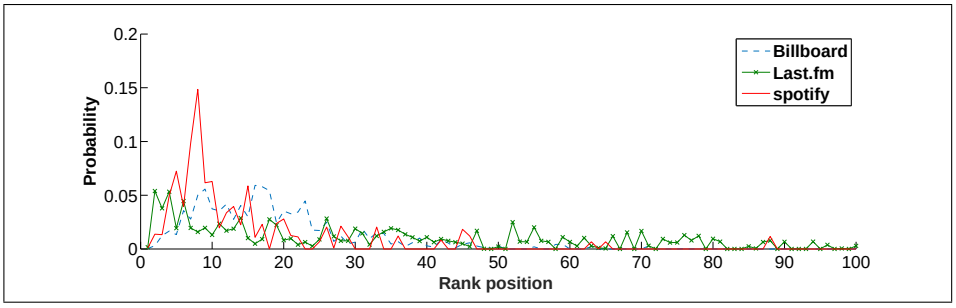

Fig. 2. Popularity chart entry position probability density.

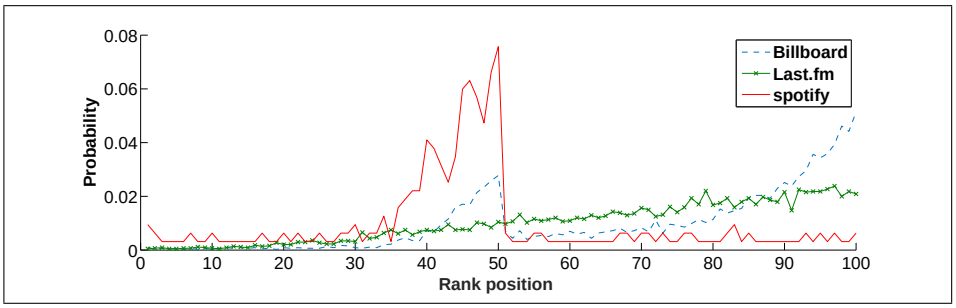

Fig. 3. Popularity chart leave position probability density.

3.3 Format and Usage

The dataset is divided into two separate parts: part A includes the relations/metadata of the tracks and their popularity while part B contains the files of the three feature-sets.

The first part is in the form of a relational database, the compact schema of which is shown in Figure 4. The archive of part A contains the SQL statements that will create the TPD database and tables and subsequently load all the information into the tables of an existing MySQL installation. Moreover, the contents of the first part are also provided in CSV format, in order to support fast use of the data and alleviate the necessity for a relational database. The second part consists of compressed archives of bz2 type that contain the feature-sets in a one file with features per track manner.

The complete TPD can be downloaded from `http://mir.ilsp.gr/track_popularity.html`.

4 Future Direction of the Dataset

The TPD is not without issues that can be ameliorated in future versions. One of these issues pertains to the automatic selection of album including each popular track: as more than one such albums may exist (hit collections, re-publication of

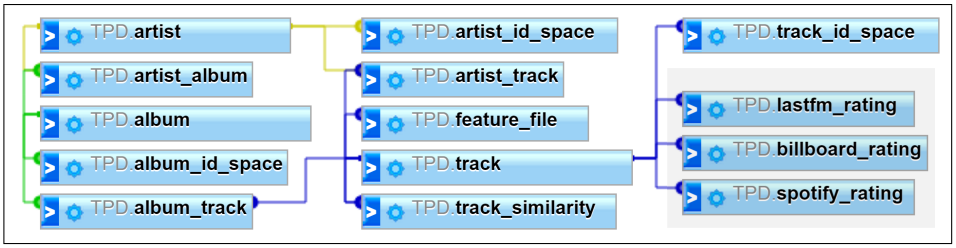

Fig. 4. Schema for the metadata and the popularity of the tracks.

the same artist, etc), there is no easy way to select the appropriate other than manual filtering. Moreover, the requirement of having access to the content of both popular and non-popular tracks elevated the complexity and timely conclusion of the collection process, which in order to remain within limits affected the size of the popularity records collected from the only source, Billboard, containing information not included in the TPD.

Some of the future actions that would greatly ameliorate the TPD are:

API A documented API for the purposes of accessing from a single point, aggregated, integrated and fully up-to-date popularity information.

Automated updates The design and implementation of a fully automated collection and integration web-based service that will update the dataset by harvesting the sources using event-driven or periodical triggers.

Popularity sources The addition of more popularity sources mostly oriented to social networks, such as twitter based hash-tags (e.g. *#nowplaying* with mention of track's metadata) as well as directly collecting tracks' airtime from e-radios using common protocols (e.g. *Shoutcast*, *Icecast*, etc).

5 Conclusion

This work introduces the Track Popularity Dataset. The dataset is, to the best of the authors' knowledge, the first complete attempt to create an integrated dataset for the purposes of mining information from musical track popularity. It includes three different sources of popularity definition with records ranging from 2004 to 2014, a mapping between different track/ author/ album identification spaces, information for the remaining, non popular, tracks of an album with popular track(s), contextual similarity between tracks and ready for MIR use extracted features.

Despite the inherent difficulty of popularity prediction prior to or during the initial period of a track's release, such a process has long been a requirement of the musical industry, while interestingly enough, the gains of such a prediction also profit artists and listeners. Thus, the availability of datasets that will allow music information researchers to experiment and compare their methods would greatly support the advancement of the research direction.

Future directions of the dataset include its manual filtering in order to enhance its content, the creation of an API for the dissemination of the dataset's information, an automated collection of up-to-date popularity information process and the expansion of the sources by addition of social networks and e-radios.

References

1. Bischoff, K., Firan, C.S., Georgescu, M., Nejdl, W., Paiu, R.: Social knowledge-driven music hit prediction. In: International Conference on Advanced Data Mining and Applications. pp. 43–54 (2009)
2. Burgoyne, J.A., Bountouridis, D., Balen, J.V., Honing, H.: Hooked: A game for discovering what makes music catchy. In: International Society for Music Information Retrieval Conference. pp. 245–250 (2013)
3. Chon, S.H., Slaney, M., Berger, J.: Predicting success from music sales data: A statistical and adaptive approach. In: ACM Workshop on Audio and Music Computing Multimedia. pp. 83–88 (2006)
4. Dhanaraj, R., Logan, B.: Automatic prediction of hit songs. In: International Society for Music Information Retrieval Conference. pp. 488–491 (2005)
5. Gkiokas, A., Katsouros, V., Carayannis, G., Stajylakis, T.: Music tempo estimation and beat tracking by applying source separation and metrical relations. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 421–424 (2012)
6. Karydi, D., Karydis, I., Deliyannis, I.: Legal issues in using musical content from itunes and youtube for music information retrieval. In: International Conference on Information Law (2012)
7. Kim, Y., Suh, B., Lee, K.: #nowplaying the future billboard: Mining music listening behaviors of twitter users for hit song prediction. In: International Workshop on Social Media Retrieval and Analysis. pp. 51–56 (2014)
8. Koenigstein, N., Shavitt, Y., Zilberman, N.: Predicting billboard success using data-mining in p2p networks. In: International Symposium on Multimedia. pp. 466–470 (2009)
9. Makris, D., Kermanidis, K., Karydis, I.: The greek audio dataset. In: Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology, vol. 437, pp. 165–173. Springer Berlin Heidelberg (2014)
10. McEnnis, D., McKay, C., Fujinaga, I., Depalle, P.: jaudio: A feature extraction library. In: International Society for Music Information Retrieval Conference. pp. 600–603 (2005)
11. Ni, Y., Santos-Rodriguez, R., McVicar, M., De Bie, T.: Hit song science once again a science? In: International Workshop on Machine Learning and Music. ACM (2011)
12. Pachet, F.: Hit song science. In: Tao, T..O. (ed.) Music Data Mining, chap. 10, pp. 305–326. Chapman & Hall / CRC Press (2011)
13. Pachet, F., Roy, P.: Hit song science is not yet a science. In: International Society for Music Information Retrieval Conference. pp. 355–360 (2008)
14. Singhi, A., Brown., D.G.: Hit song detection using lyric features alone. In: International Society for Music Information Retrieval Conference (2014)