The Million Playlists Songs Dataset: a descriptive study over multiple sources of user-curated playlists

Felipe Falcão[1*], **Daniel Mélo**[1†]

[1]Laboratory of Distributed Systems – Federal University of Campina Grande
Av. Aprígio Veloso, s/n, Bloco CO – 58.429-900, Campina Grande, PB

felipev@lsd.ufcg.edu.br, danielgondim@lsd.ufcg.edu.br

Abstract

User interest for playlists is increasing as current music streaming services become more and more popular. In order to get sets of songs that best match current musical needs (e.g. size, diversity, mood), one has to select a compatible playlists source between a representative number of options. Most of available music streaming platforms (e.g. Spotify, Pandora, Deezer) already contain playlists searching mechanisms, but as a secondary source of such information we have websites that allow users to submit, manage and publish their own playlists, organizing them according to some specific criteria. This paper proposes a descriptive study over four of these websites in such way that it categorize the groups of playlists available on each one. By recursively crawling and querying data from these sources and enriching it with high-level acoustic information fetched from AcousticBrainz, we were able to build a dataset called *Million Playlists Songs Dataset* which guided the descriptive process and is now available for further investigation.

1. Introduction

Nowadays, with the spread of music streaming services such as Spotify [1], Pandora [2], Google Play Music [3], Deezer [4] and etc., where it is possible to find millions of songs quickly and easily, it is quite common for users to organize their music creating or searching for playlists that suit their current musical needs (mood, size, genre, artist, diversity, etc.).

For being such a common concept nowadays in music, playlists have become an important object of study for the Music Information Retrieval (MIR) area. One of the great challenges involved in this study is to understand the user behavior when creating playlists. Is there a preservation in the songs features, or they prefer a heterogeneity? Is there any change in the choice of songs according to the context in which the playlist was created? As can be seen, many variables are involved in this activity, making this analysis quite complex for the MIR area.

In order to help resolve these questionings, this paper proposes a brief descriptive analysis of four playlists data sources. These sources are websites that allow users to create, manage, and share playlists manually. These sites are: 8tracks, Art of the Mix, Playlists.net and *Vagalume* (a description of these sites can be found in section 3.1). This analysis seeks to categorize groups of playlists, identifying common characteristics that may indicate user preferences.

And as a second contribution, this work also establishes the creation of a new data source, composed of all the data of the four websites analyzed, but also enriched with high-level acoustic characteristics of the songs. Such information could be retrieved using the MusicBrainz [5] and AcousticBrainz [6] platforms. Thus, as far as we know, we provide the community with a totally innovative dataset, containing not only songs from playlists, but also their high-level acoustic features.

This paper is structured as follows: in section 2 we have identified some related works, and also emphasize the novelties of this work. Section 3

*Supported by CAPES.
†Supported by CAPES.

[1]https://www.spotify.com/
[2]http://www.pandora.com
[3]https://play.google.com/music
[4]https://www.deezer.com
[5]https://musicbrainz.org/
[6]https://acousticbrainz.org/

describes how the dataset used in this work was developed. All the descriptive analysis of the dataset can be found in section 4. And finally, in section 5, we discuss about the conclusions of this work as well as possible future work.

2. Related Work

It is notable that with the spread of streaming music services, such as Spotify, Deezer, Pandora, etc., the amount of music available to users has increased significantly. To keep up with this increase and improve the user experience, multiple platforms provide the ability to create and share playlists

We can divide the activity of creating playlists into two large groups: (i) automatic generation and; (ii) manual. This first group has already been well explored in research in the area of MIR as reviewed by Bonnin and Jannach [1], indicating ways for automation based on user's listening habits [2], grouping songs by similarity of high level characteristics [3], user real-time physiological feedback [4], as well as similarity of their frequency spectrum [5]. The manual generation of playlists requires further investigation by researchers, since it is necessary to understand the behavior of users when creating playlists. Some steps have already been taken in this direction [6], in addition there are also works that examine corpus of playlists created manually [7], but these works are still vague.

Besides that, there are few datasets that incorporate data from playlists created manually. We can cite the Art of The Mix, made available by McFee and Lancriet [8], #nowplaying [9] and 30Music [10]. Although they are datasets with a considerable amount of data, they are restricted only to common descriptions such as album name and tags (in addition to the song and artist name).

This work fills this gap since it proposes the creation of a huge dataset, with almost 2 million musical entries of playlists, where not only the common features are stored, but also the high-level acoustic ones of such songs. These features are obtained through AcousticBrainz [11].

3. The Dataset

The *Million Playlists Songs Dataset* - MPSD (deliberately a tribute to the *Million Song Dataset* [12], whose work guided us during our efforts) comprises data fetched from four different sources of user-curated playlists. Since most of current studies [10, 13] already considered monitoring broadly used platforms such as Twitter, Last.FM and Spotify, we have, on the other hand, focused on crawling data from secondary platforms which, although not holding as much users as the aforementioned systems, also remains as an unexplored source of playlists data, with a representative number of enthusiasts and curated playlists.

3.1. Data Sources

The methodology applied during the preparation of our dataset is hybrid and featured by both crawling and querying approaches. The choice between some of these approaches was defined by how the desired data was structured on their websites and how easy it would be for authors to have the maximum of data available for analysis in the shortest amount of time.

The first source of playlists data included on the dataset is the *Vagalume* website. *Vagalume* [7] is a music portal created in Brazil on 2002, initially conceived as a public database of song's lyrics. As years went by and the platform received more attention by the community (specially from Brazil and Portugal), features were expanded and users were then allowed to upload public content, such as public playlists composed by Youtube music videos. All the dataset *Vagalume*-related data was fetched by crawling playlists pages from the main *Vagalume* profile [8] and all playlists from his respective followers, totalizing 35,600 profiles. Profiles without registered playlists were ignored.

The proposed dataset also comprises data coming from the playlists.net [9] website. By also crawling playlists pages from this platform we have enriched our dataset with playlists hosted on Spotify and submitted to this platform by

[7] https://www.vagalume.com.br
[8] https://meu.vagalume.com.br/sitevagalume
[9] http://playlists.net/

a very active community of users interested on discovering playlists that sometimes cannot be found directly on Spotify browser. Althought site creators claim to have about 170,000 registered playlists, many of them are not available on Spotify anymore and some others were not listed on the webpage at crawling time.

A well-known platform was used as data source for our work: 8tracks.com [10]. Founded in 2006, 8tracks is a collaborative platform that allows users to share and discover music in a simple, legal, and free way. On this platform, users can create and share playlists with at least eight songs. The data from this platform was provided to us directly by its administrator.

Finally, we also use the data from the Art of the Mix website [11]. This site integrates playlists created on iTunes, nightly. These data were provided by McFee and Lancriet [8] and contains information from more than 100,000 playlists.

3.2. Crawling

In order to automate the process of recursively looking into multiple sections of several websites, the process of building MPSD was aided by some computer-aided software engineering (CASE) tools that provided us some ready-to-use features without which any of the now-available artifacts would be published in time.

Crawlings were fully accomplished by running Python scripts along with Scrappy [12] tasks that recursively visited thousands of webpages to fetch desirable data stored on nested tags of HTML documents. Scrappy is a fast high-level web crawling open source framework written in Python to assist developers on tasks based on the extraction of structured data from websites and APIs. By providing mechanisms of recursive crawlings, this tool allows users to start looking at specified URLs, extract desired information present on the HTML document and search for external links this page might have to proceed with the crawling loop as deep as planned. The result of this process is a list of visited pages as well as the set of extracted data.

Besides, to simulate user-specific behavior (link clicks, in this case) scripts were also enriched with Selenium [13] features. Even though originally designed for software-testing tasks, Selenium suite was able to provide us some web-browser automation tools that helped us to load some data that could only be available by interacting with web interface elements via link clicks, since Scrappy isn't currently able to perform such kind of operation.

All these aforementioned tools were combined in order to create a powerful and generic web crawler that initially ran over the two chosen web-based sources (*Vagalume* and Playlists.net) at the same time in a single computer [14] to produce the expected outputs. These tasks took about eight uninterrupted days to be finished using our available infrastructure.

3.3. Extraction

Due to the large amount of data and the limitation of time and resources for processing, a reduction in the amount of data used from 8tracks and Art of the Mix was required.

For 8tracks, we only extracted from the database playlists with more than ten songs. Due to inconsistencies in the given database, several playlists with fewer than ten songs were returned, as their *tracks_count* attribute indicated a value greater than 10, but had a smaller list of songs. As this inconsistency would not cause any damage to our work, we consider all the data returned in this extraction, which counted a total of 27,606 playlists, dated from September 2007 until June 2012.

32,681 playlists were extracted from Art of the Mix. This value was reached after we left an extraction script running for 7 uninterrupted days with the resources we had. These playlists cover the period from January 1998 to June 2011.

3.4. Acoustic Enrichment

As our crawling and querying techniques were extracting playlists and tracks metadata

[10] https://8tracks.com/

[11] http://www.artofthemix.org/

[12] https://scrapy.org

[13] http://www.seleniumhq.org

[14] In our experiments, we used an 8-core Intel(R) Core(TM) i5-4200U CPU @ 1.60GHz with 8GB of RAM memory running Ubuntu 16.04.2 LTS. All our codes were implemented and executed using Python 2.7.

available on all of the chosen sources, we also tried to enrich even more the gathered information by appending to it some extra high-level acoustic features available for querying on AcousticBrainz database. For so, every single track on each of the studied sources was queried on MusicBrainz so it could have assigned to itself a MusicBrainz Identifier (MBID) which would be used to fetch acoustic data on AcousticBrainz, if available.

Since AcousticBrainz is a recent platform and also taking into consideration that the queried songs came from secondary sources we could not get all the information planned. Instead, we realized that only 10,45% of our comprised songs were able to be enriched with acoustic data. Even though it is less than half of all dataset songs, we consider that it is a representative result for this first effort.

4. Dataset Analysis

MPSD is currently a collection of 1,993,607 tracks of 74,996 distinct playlists songs annotated with 45 field descriptors (both statistics of each source and all field descriptors can be examined on Tables 1 and 2, respectively). Altogether, 617,242 distinct track names of 221,560 distinct artists can be fully analyzed in a 576,6MB CSV file available on the project GitHub repository [15].

Table 1: Playlists statistics

Sources	# playlists	# tracks	# artists	Maximum Playlist Size
AotM	32,681	296,344	110,154	60
8tracks	27,606	115,660	50,354	226
Vagalume	9,584	124,627	6,611	2995
Playlists.net	5,125	185,291	92,236	7287

Moving forward with our analysis and deeply looking into our data we could elaborate some hypothesis about the size of our available playlists. Table 1 shows us that both *Vagalume* and Playlists.net have a smaller amount of distinct playlists stored on database, but the total crawled songs for each source (646,070 and

[15] https://github.com/felipevieira/computacao-e-musica-lsd/tree/master/sbcm-2017

Table 2: All song's field descriptors comprised by MPSD

Field	Description
source	Source that hosts the playlist (Possible values: *Vagalume*, AoTM, 8tracks, playlists.net)
user_id	An unique playlist identifier (format varies from source to source)
track_name	Song title
artist_name	Artist or band that performs that specific version of a song
mbids	List of MusicBrainz identifiers (a 36 character Universally Unique Identifier that is permanently assigned to each entity in the database)
playlist_id	An unique playlist identifier (format varies from source to source)
tags	List of labels attached to each song in order to provide extra-information about it (format varies from source to source)
playlist_name	Playlist title
danceability_value/prob	Danceability value and probability as defined by the Essentia classifier model [14] (Possible values: danceable, not_danceable)
gender_value/prob	Gender value and probability as defined by the Essentia classifier model [14] (Possible values: male, female)
genre_[dataset]_value/prob	Genre value and probability as defined by the Essentia classifier model [14] for four different datasets
ismir04_rhythm_value/prob	Rhythm value and probability as defined by Goiiyon classifier model [15]
mood_[type]_value/prob	Mood value and probability as defined by the Essentia classifier model [14] for eight different mood types
timbre_value/prob	Timbre value and probability as defined by the Essentia classifier model [14] (Possible values: bright, dark)
tonal_atonal_value/prob	Tonal/Atonal value and probability as defined by the Essentia classifier model [14] (Possible values: tonal, atonal)
voice_instrumental_prob/source	Voice/Instrumental value and probability as defined by the Essentia classifier model [14] (Possible values: voice, instrumental)

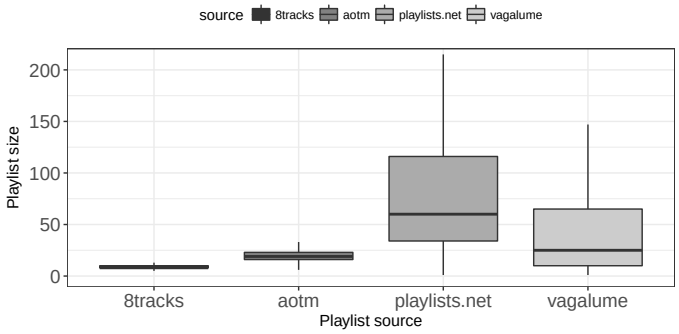

Figure 1: Playlists size distribution

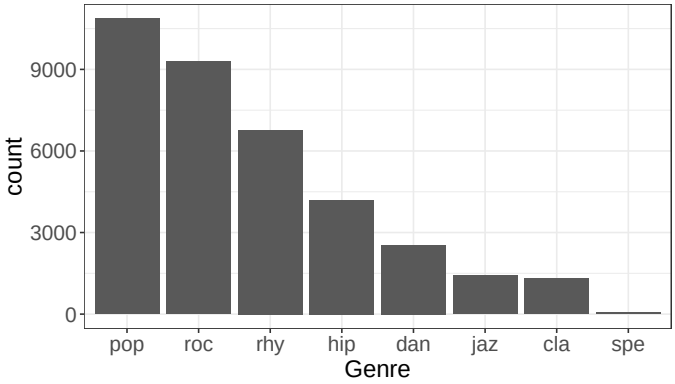

Figure 2: Genre Histogram of *Vagalume*'s playlists

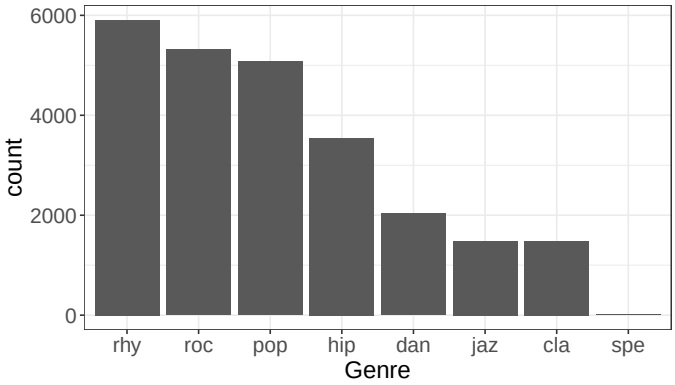

Figure 3: Genre Histogram of Playlists.net's playlists

432,351 tracks for these two sources, respectively, against 643,349 and 258,727 from 8tracks and Art of The Mix) does not reflect this minority. By checking average and standard deviation information for grouped data (Table 3) in addition with the playlists size distribution on Figure 1 we confirm our theory, concluding that even though we were able to crawl more playlists from AotM and 8tracks, the ones obtained from *Vagalume* and Playlists.net had more songs on them.

Since genre can be an important factor considered by users when searching for playlists, this study also tried to find out and understand the distribution of genres along the tracks fetched on each of our sources. Such task was aided by AcousticBrainz information added to our dataset on the Acoustic Enrichment Phase mentioned on subsection 3.4. The field used to best summarize genre information about a song was the *genre_rosamerica_value*, which estimates a song genre by using the Rosamerica Collection [16, 17] while training a classifier model that assigns one of the eight possible genre values (rhythm & blues, rock, pop, hip-hop, dance, jazz, classic and speech) to new song entries.

Our analysis shows some concrete differences between genres distributions over all four sources. While 8tracks and AotM users seems to have very similar affinities with rock and r&b (in the same order), *Vagalume* and Playlists.net users (Figures 2 and 3, respectively) prefer pop and r&b songs, respectively.

In an effort to exemplify some useful applications of MPSD on what regards a better understanding of how users behaviour when creat-

Table 3: Statistical data about playlists size

Sources	Average	Standard Deviation
AotM	19.68 tracks	6.22
8tracks	9.37 tracks	4.53
Vagalume	67.41 tracks	148.12
Playlists.net	84.13 tracks	96.78
Total	29.35 tracks	66.68

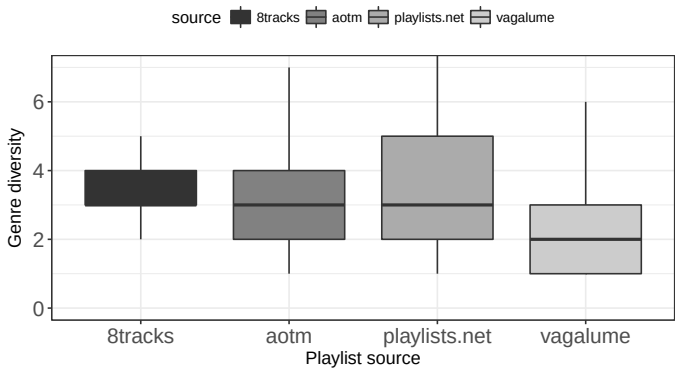

Figure 4: Playlists genre diversity distribution

ing playlists differ from one source to another, we have used some of the data available on the dataset to extract some genre-diversity insights observed on the crawled playlists. For this, we summarized our dataset to check how many distinct genres were present in each of our comprised playlists. Since Acoustic Enrichment Phase wasn't able to fetch high-level acoustic data for all playlist songs, in order to minimize the effect of this lack of data we have filtered our dataset to only consider playlists with more than five songs contemplated with acoustic data (i.e. *genre_rosamerica_value* field descriptor). The result of our analysis can be examined on Figure 4 concluding that *Vagalume* is the source with the less genre-diversified playlists (with a median of two different genres per playlist), while playlists.net sets of songs can be considered as the most diverse in terms of genre, even though its diversity median was the same as 8tracks and AoTM.

5. Further Work

In the course of this paper we were able to present all the methodology applied on the composition of the Million Playlists Songs Dataset: a now-public dataset of metadata information regarding playlists songs from four web platforms designed to allow user curation over playlists. Besides, as a way of exemplify studies that may be conducted over this set of data, a simple descriptive analysis was performed over all data to extract insights about the distinct sources of data considered when building the dataset.

There is plenty of contributions to be performed in order to complement this preliminary study. One of them is the gradual increment of this dataset with all kinds of sources and information that could not be contemplated by this current research by time and computational reasons. By conducting a long-term research, one can increment this dataset with new sources and crawl older data information, and these would be good approaches to attach even more value to this dataset.

In addition, a more detailed study of the data in this dataset could be carried out in order to extract more information about the behavior of the users during the process of creating playlists, e.g. identify the features that matter most to the users, the trends of the users, etc.

Another gap found during the current study is the lack of a complete state-of-art framework designed to extract acoustic features from as much songs as possible. Even with the arise of AcousticBrainz (with almost 5 million registered songs) we still faced minor issues when trying to fetch acoustic data from MPSD tracks, specially on what refers to brazilian music.

References

[1] Geoffray Bonnin and Dietmar Jannach. Automated generation of music playlists: Survey and experiments. *ACM Comput. Surv.*, 47(2):26:1–26:35, November 2014.

[2] Andreja Andric and Goffredo Haus. Automatic playlist generation based on tracking user's listening habits. *Multimedia Tools and Applications*, 29(2):127–151, 2006.

[3] Steffen Pauws and Berry Eggen. Pats: Realization and user evaluation of an automatic playlist generator. In *ISMIR*, 2002.

[4] Nuria Oliver and Lucas Kreger-Stickles. Papa: Physiology and purpose-aware automatic playlist generation. In *ISMIR*, volume 2006, page 7th, 2006.

[5] Beth Logan. Content-based playlist generation: Exploratory experiments. In *ISMIR*, 2002.

[6] Sally Jo Cunningham, David Bainbridge, and Annette Falconer. "more of an art

than a science": Supporting the creation of playlists and mixes. 2006.

[7] Dietmar Jannach, Iman Kamehkhosh, and Geoffray Bonnin. Analyzing the characteristics of shared playlists for music recommendation. In *RSWeb@ RecSys*, 2014.

[8] Brian McFee and Gert R. G. Lanckriet. Hypergraph models of playlist dialects. In *ISMIR*, 2012.

[9] Martin Pichl, Eva Zangerle, and Günther Specht. Towards a context-aware music recommendation approach: What is hidden in the playlist name? In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, pages 1360–1365. IEEE, 2015.

[10] Roberto Turrin, Massimo Quadrana, Andrea Condorelli, Roberto Pagano, and Paolo Cremonesi. 30music listening and playlists dataset. In *RecSys Posters*, 2015.

[11] Alastair Porter, Dmitry Bogdanov, Robert Kaye, Roman Tsukanov, and Xavier Serra. Acousticbrainz: a community platform for gathering music information obtained from audio. In *International Society for Music Information Retrieval Conference (ISMIR'15)*, 2015.

[12] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, volume 2, page 10, 2011.

[13] Eva Zangerle, Martin Pichl, Wolfgang Gassler, and Günther Specht. # nowplaying music dataset: Extracting listening behavior from twitter. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, pages 21–26. ACM, 2014.

[14] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R Zapata, Xavier Serra, et al. Essentia: An audio analysis library for music information retrieval. In *ISMIR*, pages 493–498, 2013.

[15] Fabien Goiiyon. Dance music classification: A tempo-based approach. 2004.

[16] Dmitry Bogdanov, Alastair Porter, Perfecto Herrera, and Xavier Serra. Cross-collection evaluation for music classification tasks. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2016.

[17] Enric Guaus i Termens. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.* PhD thesis, Universitat Pompeu Fabra, Barcelona Barcelona, 2009.