-

Análisis de Similitud entre Playlists del Million Playlist Dataset: Un Enfoque de Minería de Datos Basado en Atributos Musicales y Técnicas de Clustering

Paula D. Velosa, Johan A. Díaz, y Fabián L. Lopez

[1]

***Resumen*—Este trabajo presenta el desarrollo de un pipeline de minería de datos orientado al análisis de similitud entre playlists utilizando el challenge set del Million Playlist Dataset (MPD) de Spotify. Con el fin de enriquecer los datos y capturar una representación más completa de las características musicales, se integró información adicional proveniente de la Spotify Web API, MusicBrainz y AcousticBrainz. El proyecto sigue el enfoque completo del proceso KDD (Knowledge Discovery in Databases), incluyendo la integración y limpieza de datos, un análisis exploratorio detallado y la aplicación de técnicas de preprocesamiento orientadas a la reducción de dimensionalidad y la normalización de atributos. Posteriormente, se implementan métodos de agrupamiento (clustering) y métricas de similitud para identificar patrones latentes y estructuras de afinidad entre playlists. Los atributos considerados abarcan tanto propiedades acústicas (como danceability, energy, valence y tempo) como metadatos relevantes (géneros, popularidad, artistas), permitiendo una caracterización más rica de los gustos musicales. Este enfoque busca aportar herramientas para la comprensión de preferencias colectivas, con aplicaciones potenciales en sistemas de recomendación y análisis de comportamiento musical.**

***Index Terms*—AcousticBrainz, análisis exploratorio, artistas, características acústicas, clustering, dimensionalidad, energía, géneros, KDD, Million Playlist Dataset, MusicBrainz, normalización, popularidad, preprocesamiento, recomendación, similitud, Spotify,**

I. INTRODUCCIÓN

En la era del consumo digital de música, las plataformas de streaming como Spotify han transformado la forma en que los usuarios descubren y organizan sus preferencias musicales. En este contexto, las playlists no solo representan colecciones arbitrarias de canciones, sino también manifestaciones estructuradas de gustos individuales y colectivos. Comprender las relaciones entre playlists, especialmente en términos de similitud, resulta fundamental para el diseño de sistemas de recomendación más eficientes y personalizados.

Este trabajo se enmarca en el análisis del challenge set del Million Playlist Dataset (MPD), un subconjunto de 10.000 playlists parciales provisto por Spotify para tareas de predicción y recuperación de contenido musical. Para enriquecer este conjunto de datos y obtener una representación más completa de cada ítem musical, se incorporó información proveniente de tres fuentes externas: la Spotify Web API, MusicBrainz y AcousticBrainz. Esta fusión de datos permitió acceder a atributos acústicos (como danceability, energy, valence y tempo), metadatos contextuales (como género, artista, popularidad) y descriptores semánticos adicionales.

El enfoque metodológico adoptado sigue el ciclo completo del proceso de descubrimiento de conocimiento en bases de datos (KDD), abarcando desde el análisis exploratorio y el preprocesamiento hasta la aplicación de técnicas de agrupamiento (clustering) y medición de similitud. El objetivo principal es identificar patrones latentes y agrupaciones naturales de playlists que compartan afinidades musicales, aportando así al entendimiento de las preferencias de los usuarios y a la mejora de sistemas de recomendación en entornos musicales.

II. DESCRIPCIÓN DE LOS DATOS

El conjunto de datos a considerar es un subconjunto de 10.000 playlist parciales provenientes del challenge set del Million Playlist Dataset (MPD) de Spotify. Cada registro incluye detalles como el título de la playlist, el título de las pistas y metadatos adicionales como la última fecha de edición y el número de playlists editadas.

Los datos de este dataset abarcan playlists públicas creadas por usuarios estadounidenses desde enero de 2010 hasta noviembre de 2017. Cada playlist contiene 6 variables y una lista de hasta 100 canciones, de las cuales tienen cada una 8 atributos.

TABLE I
DESCRIPCIÓN DE VARIABLES NIVEL PLAYLIST

Variable	Tipo	Descripción
pid	Entero	ID de la playlist
name	Texto	Nombre de la playlist (opcional)
num_holdouts	Entero	Canciones ocultas
num_samples	Entero	Canciones visibles
num_tracks	Entero	Total de canciones
tracks	Lista	Lista de canciones visibles

Adicionalmente, para enriquecer el dataset se agregan datos provenientes de MusicBrainz y AcousticBrainz, donde hay

[1]None

-

información adicional sobre las canciones en forma de descripciones acústicas y metadatos generados automáticamente a partir del análisis de las señales musicales. Estos serán atributos adicionales que buscan capturar elementos musicales como género, estado de ánimo, ritmo y timbre.. A cada canción registrada en los playlist del conjunto de datos original se le asocian los nuevos datos, que consisten en 22 variables adicionales. El dataset puede utilizarse para tareas de minería de datos como clasificación de géneros, detección de estados de ánimo, análisis de estilos musicales y segmentación de audiencias.

TABLE II
DESCRIPCIÓN DE VARIABLES NIVEL CANCIÓN (DATASET ORIGINAL DEL CHALLENGE)

Variable	**Tipo**	**Rango estimado**	**Descripción**
pos	Entero	0 - 99	Posición en la playlist
track_name	Texto	---	Nombre de la canción
track_uri	Texto	22 caracteres	URI único en Spotify
artist_name	Texto	---	Artista principal
artist_uri	Texto	22 caracteres	URI del artista
album_name	Texto	---	Nombre del álbum
album_uri	Texto	22 caracteres	URI del álbum
duration_ms	Entero	30.000 - 600.000 ms	Duración de la canción

TABLE III
DESCRIPCIÓN DE VARIABLES NIVEL CANCIÓN (MUSICBRAINZ)

Variable	**Tipo**	**Descripción**
mbid	Texto	MusicBrainz ID único para la canción
genre_mb	Texto	Género según MusicBrainz (puede ser nulo)
bpm	Real	Tempo promedio en beats por minuto
energy	Real	Intensidad percibida de la canción (escala relativa)
danceability_ll	Real	Medida estimada de bailabilidad basada en modelo de regresión logística
loudness	Real	Nivel relativo de volumen
rating_value	Entero	Calificación media (nula en muchos casos)
rating_votes	Entero	Número de votos asociados (nulo en muchos casos)

TABLE IV
DESCRIPCIÓN DE VARIABLES NIVEL CANCIÓN (ACOUSTICBRAINZ, HIGH LEVEL)

Atributo	**Tipo**	**Valores posibles**	**Descripción breve**
danceability	Categórico	danceable, not_danceable	Facilidad para ser bailada
gender	Categórico	female, male	Género vocal predominante
genre_*	Categórico	pop, rock, electronic, etc.	Estimaciones por modelos distintos de clasificación de géneros
ismir04_rhythm	Categórico	Samba, Rumba, Tango, etc.	Estilo rítmico según patrones de danza
mood_*	Categórico	happy, sad, party, relaxed, etc.	Afecto emocional predominante
moods_mirex	Categórico	Cluster1 - Cluster5	Agrupación emocional según MIREX
timbre	Categórico	bright, dark	Color sonoro o textura de la canción
tonal_atonal	Categórico	tonal, atonal	Presencia o ausencia de una tonalidad definida
voice_instrumental	Categórico	voice, instrumental	Predominio vocal o instrumental

Los datos de los atributos expuestos en la tabla IV están disponibles en forma de una matriz de probabilidad: cada atributo tiene un diccionario donde se listan las probabilidades de cada categoría.

Frente a todo el conjunto extendido de datos que se tiene se puede decir que se tiene una enorme dimensionalidad: 23 atributos por canción y 6 por playlist. Incluso se puede llegar a considerar variable ya que cada playlist puede albergar un número distinto de canciones. La dispersión en este conjunto de datos es considerablemente baja, ya que las playlist son altamente únicas. Únicamente se podría decir respecto a la dispersión que hay algunas canciones que pueden ser compartidas por varias playlists. La resolución de este conjunto es excepcionalmente alta, teniendo datos al nivel de cancion, y algunos de estos al nivel de los reales.

III. OBJETIVO DE MINERÍA DE DATOS

El objetivo general de aplicar minería de datos sobre este conjunto extendido es descubrir agrupaciones y patrones de similitud musical entre playlists del challenge set. Entre los objetivos específicos encontramos:

- Filtrar playlists con al menos una canción visible.
- Comparar playlists entre sí usando artistas, canciones y nombres.
- Agrupar playlists similares (clustering).
- Visualizar y analizar temáticas o estilos musicales implícitos.