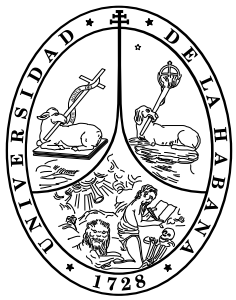
Facultad de Matemática y Computación

Universidad de La Habana



Tesis de diploma de la carrera Ciencia de la Computación

Recuperación semántica de música utilizando embeddings y modelos de clasificación.

Autor: **Niley González Ferrales**

Tutor: **Dr. Yudivian Almeida**

La Habana

2023

A la ciencia.

Opinión del tutor

Dr. Yudivian Almeida
Facultad de Matemática y Computación
Universidad de la Habana
Enero, 2023

Resumen

Palabras claves:

Abstract

Keywords:

Índice general

Introducción

El auge de la tecnología digital, en los últimos años, ha llevado a una explosión en la cantidad de contenido multimedia accesible en línea. Incluyendo texto (libros, páginas web, ...), imágenes, videos y música; la red se ha convertido en un repositorio de información con un crecimiento vertiginoso [1]. La necesidad de buscar en bases de datos cada vez más grandes ha impulsado importantes avances en el campo de la búsqueda y recuperación de información. Existen motores de búsqueda para diversos contenidos como páginas webs, imágenes y videos; sin embargo la música no es accesible de la misma forma [2]. En este contexto, la presente tesis se propone explorar nuevas técnicas y enfoques para la búsqueda y recuperación de música en grandes bases de datos, con el objetivo de mejorar la accesibilidad de este recurso.

0.1. Contexto histórico/social

Durante las últimas décadas la tecnología computacional se ha desarrollado al punto de estar presente, de una manera u otra, en casi todos los procesos cotidianos de los seres humanos. La velocidad a la que se generan grandes volúmenes de datos (conocido comunmente como *big data*), en la actualidad, desafía las capacidades de procesamiento computacional [3].
Una característica fundamental de estos volúmenes de datos es la gran variedad que presentan, siendo alrededor del 80% de naturaleza no estructurada. Y precisamente, la recuperación de información se define como encontrar resultados de naturaleza no estructurada que satisfaga una necesidad de información dentro de una gran colección de datos [4].
Los sistemas de recuperación de información (SRI) consisten en tecnologías y métodos diseñados para la búsqueda, almacenamiento, recuperación y organización de información. Estos sistemas son esenciales en la gestión de

grandes cantidades de datos en diversos entornos, como bibliotecas digitales, bases de datos en línea y motores de búsqueda en la web.
La recuperación de información utilizando computadoras se remonta a la década de 1950, y desde entonces se han desarrollado grandes ideas en el campo. Entre estas ideas se encuentran la creación de rankings de documentos, la representación vectorial de documentos y consultas, el agrupamiento de documentos similares, la asociación de términos con similitudes semánticas, la introducción de la frecuencia inversa en documentos y los modelos de semántica latente.[5]
Uno de los avances recientes en la recuperación de información es el uso de redes neuronales. Este es un campo que ha avanzado a pasos agigantados en la última década, en gran parte gracias al aumento del poder computacional introducido con las GPU. Los modelos de lenguaje, como Word2Vec [6], GloVe [7] y los más recientes, basados en transformers, como BERT [8] y GPT [9], han demostrado una capacidad excepcional para capturar el significado semántico y la relación contextual entre palabras y frases. Estos modelos, operan con representaciones de *embeddings* a partir de diferentes enfoques cada uno.
Los embeddings son representaciones, en un espacio de relativamente baja dimensionalidad, de tokens como frases, párrafos o documentos, a partir de un espacio vectorial de alta dimensionalidad, donde cada dimensión corresponde a una característica o atributo del lenguaje aprendido. Al aplicar estos *embeddings* a la representación de documentos y consultas, se ha logrado una mejora significativa en la precisión y la relevancia de los resultados de recuperación de información [10].

0.2. Motivación

A pesar de que los SRI se han centrado principalmente en la recuperación de información con forma textual, se ha reconocido la necesidad de adaptar el entorno para manejar todo tipo de datos (como imágenes, videos, audios) [11]. La recuperación de información en datos no textuales presenta desafíos únicos, como la necesidad de comprender el contenido visual o auditivo de los datos. En el caso de imágenes, por ejemplo, la búsqueda puede basarse en características visuales como el color, la textura o la forma. En el caso de videos y audios, la búsqueda puede basarse en características como el tono, el ritmo o el contenido verbal. En los últimos años, ha habido avances significativos en la recuperación de información

en datos no textuales, impulsados por el desarrollo de algoritmos de *deep learning* y el aumento de la capacidad computacional. Estos avances han permitido el desarrollo de sistemas más precisos y eficientes para la recuperación de imágenes, videos y audios.
Una herramienta utilizada comúnmente para recuperar información multimedia es representar los datos como texto, empleando metadatos [12]. Estos metadatos sirven para describir y normalizar la representación de la información digital, lo que a su vez facilita su posterior búsqueda y recuperación. A pesar de sus beneficios, el uso de metadatos fijos conlleva limitaciones significativas como la imposibilidad de adaptarse a cambios en la información y la falta de precisión en la descripción de la información, así como la restricción a una cantidad limitada de información descriptiva. En este sentido, los metadatos fijos pueden resultar insuficientes para abarcar la complejidad inherente a la información multimedia, lo que a su vez puede dificultar su precisa recuperación.
La creación de metadatos suele requerir un enfoque manual, o en su defecto, la utilización de modelos de inteligencia artificial. Sin embargo, estos modelos, al depender de los datos con los que fueron entrenados, sufren dificultades para generalizar adecuadamente para contextos específicos. Además, la naturaleza dinámica de la información multimedia puede llevar a que los metadatos fijos se vuelvan obsoletos con el tiempo, lo que impacta negativamente la precisión y exhaustividad en la recuperación de información.
Dadas las limitaciones que presenta la recuperación utilizando metadatos, se han estudiado otras alternativas. Un ejemplo es la consulta por tarareo (*query-by-humming*) o la consulta por ejemplo (*query-by-example*), en la que la consulta se representa mediante una grabación de audio y el sistema recupera la canción correcta (la más similar). La recuperación basada en contenido consiste en buscar audio que coincida con una consulta de audio. Dado un audio de ejemplo, se devuelven los audios más similares en la base de datos. Sin embargo, satisfacer los requisitos para introducir un audio puede resultar difícil para los usuarios. Por otro lado, las consultas detalladas en lenguaje natural con forma libre, constituyen una interfaz familiar ampliamente utilizada en los motores de búsqueda actuales, lo que permite solicitar música con información semántica.
Permitir a los usuarios buscar música utilizando oraciones en lenguaje natural introduce una capa adicional de complejidad, ya que requiere el puente entre la semántica del lenguaje y la representación de la música en sí misma.
Uno de los desafíos clave, en la integración del lenguaje natural y análisis

de audios en los SRI, se encuentra en la representación efectiva de la música en función de las consultas textuales. Hasta donde se ha podido comprobar, este problema no ha sido abordado utilizando las emergentes capacidades semánticas y contextuales de los grandes modelos de lenguajes (LLM por sus siglas en inglés). Tampoco se ha observado una estrategia de caracterización de la música empleando modelos pre-entrenados de inteligencia artificial, como una alternativa a entrenar modelos de *music captioning* que son la forma más directa de representar música como texto. Depender de un único modelo trae desventajas en cuanto a que las características que se pueden extraer, están limitadas a las que aparecen en los datos de entrenamiento; esas desventajas son acentuadas por el conocido problema que hay con la indisponibilidad de *datasets* de tamaño adecuado y con amplio rango de modalidades [13].

0.3. Antecedentes

Este trabajo es pionero en la investigación de recuperación semántica de música en la Facultad. No se cuentan con investigaciones en esta área.

0.4. Problemática

Durante una exahustiva investigación no se encontró ningún sistema o diseño, que permita organizar y recuperar música de una gran base de datos, utilizando consultas en lenguaje natural y con componentes semánticos; y que además sea escalable respecto a las características(o *features*) que se conocen de la música.

0.5. Objetivos

El objetivo de esta investigación es el diseño e implementación de un prototipo de una plataforma que permita realizar consultas en lenguaje natural sobre una base de datos de música.
Los objetivos específicos planteados para dar cumplimiento al objetivo general son:

- Estudiar los resultados más recientes relacionados con recuperación de información multimedia, en particular de música. Indagando en

recuperación con consultas en lenguaje natural.

- Modelar una sistema de extracción de *features* que sea flexible, permitiendo actualizar y añadir información sobre la música.
- Modelar un SRI que devuelva canciones relevantes a consultas en lenguaje natural. Intentando incluir similitud semántica sin sacrificar mucha eficiencia.
- Implementar la plataforma modelada anteriormente.
- Evaluar el Sistema de Recuperación y comparar los resultados con modelaciones similares en la literatura.
- Realizar una página web utilizando django *framework* que permita escribir consultas y reproducir la música que sea resultado de la consulta; como interfaz para probar el prototipo implementado.

0.6. Preguntas ciéntificas

A través del documento se abordarán las siguientes preguntas con la intención de resolver el problema planteado:

- ¿Un *framework* escalable conformado por modelos de *Music Information Retrieval*(MIR) de clasificación de música es, al menos, igual de efectivo al extraer características de música, que los modelos existentes de *auto-tagging* o *captioning*?
- ¿Se puede diseñar un sistema de recuperación que utilice las características extraídas (que son básicamente metadatos), pero sin las limitaciones que conllevan?
- ¿Un SRI utilizando *embeddings* es más efectivo en cuánto a similitud semántica que empleando vectores de bolsa de palabras?

0.7. Estructura de la tesis

A continuación se describe la estructura del documento. En el capítulo **Preliminares** se define el marco teórico del trabajo, que abarcará los modelos de recuperación de información, los acercamientos a recuperación de audios y música, modelos de lenguaje y como sobreponerse al *semantic gap*. En el capítulo **Propuesta** se presenta la propuesta específica para atacar el

problema planteado, mientras que en el capítulo **Detalles de implementación** se describen los detalles de implementación incluyendo las alternativas que no se pudieron comprobar. En el capítulo **Resultados computacionales** se muestran las evaluaciones del modelo de recuperación y similares para los pasos intermedios del sistema. Finalmente, en se muestran las conclusiones, recomendaciones y el camino al trabajo futuro sobre el tema.

Capítulo 1

Preliminares

Capítulo 2

Detalles de implementación

Conclusiones

Bibliografía

[1] Hannah Ritchie, Edouard Mathieu, Max Roser, and Esteban Ortiz-Ospina. Internet. *Our World in Data*, 2023. https://ourworldindata.org/internet. (Citado en la página 1).

[2] A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2021. (Citado en la página 1).

[3] Bayan Alabdullah, Natalia Beloff, and Martin White. Rise of big data – issues and challenges. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–6, 2018. (Citado en la página 1).

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. (Citado en la página 1).

[5] Carlos Fleitas Aparicio and Marcel E. Sanchez Aguilar. Conf, 2021. (Citado en la página 2).

[6] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. (Citado en la página 2).

[7] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. (Citado en la página 2).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. (Citado en la página 2).

[9] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. (Citado en la página 2).

[10] Shay Palachy Affek. Document embedding techniques, 2019. Accessed on October, 2023. (Citado en la página 2).

[11] Jian Zhang. Lecture 5: Multimedia information retrieval, 2007. (Citado en la página 2).

[12] Seungheon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. Toward universal text-to-music retrieval. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022. (Citado en la página 3).

[13] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. Multimodal music information processing and retrieval: Survey and future challenges. *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, pages 10–18, 2019. (Citado en la página 4).