Qualitative Analysis of Content-Based Music Retrieval Systems

HARALD EIBENSTEINER, k01300179
HADI SANAEI, k11733444
LUKAS TROYER, k12006666
LUKAS WAGNER, k01357626
BRANKO PAUNOVIĆ, k12046370

This qualitative analysis showcases the different strengths and weaknesses of the examined music retrieval systems. While the random baseline system served as a control test, the text-based systems, depending on the word embedding in use, showcased varied influences on similarity classification. These findings emphasize the importance of word embedding choice in music retrieval systems and contribute to a deeper understanding of their respective underlying algorithms.

CCS Concepts: • **Information systems** → *Presentation of retrieval results*; **Relevance assessment**.

Additional Key Words and Phrases: music retrieval, music similarity, retrieval evaluation

1 INTRODUCTION

This lab report aims to conduct a qualitative analysis of outputs from various content-based music retrieval systems utilizing different word embeddings. The focus lies in examining how different text-based systems, employing distinct word embeddings, namely TF-IDF, BERT, and word2vec, perform in music retrieval tasks. This analysis involves four different retrieval systems: a random baseline system and three text-based systems. The evaluation is conducted on a subset [1] of the Music4All-Onion [4] dataset, with the random baseline system serving as a control measure. The outputs were analyzed qualitatively, focusing on the relevance and accuracy of the retrieved music tracks based on the input query.

2 LAB

2.1 Setup and Approach

Four different retrieval systems were evaluated [2]:

- **Random Baseline**: This system randomly selects tracks from the song catalog.
- **Cosine similarity with TF-IDF**: Text-based system utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) [2] measure for the embeddings.
- **Cosine similarity with BERT**: Text-based system utilizing the Bidirectional Encoder Representations from Transformers (BERT) [1] model for word embeddings.
- **Cosine similarity with word2vec**: Text-based system utilizing the word2vec [3] model for the embeddings.

All text-based systems used the cosine similarity measure, which was intentionally kept constant, in order to minimize the number of moving parts in the lab environment. The cosine similarity indicates the similarity between two non-zero vectors in a multi-dimensional space, which, in our context, relates to the similarity in music tracks based on their textual data like lyrics and titles.

Each system was tested on the same subset of the Music4All-Onion dataset to ensure a fair comparison. In order to ensure reproducibility with regards to randomness in the experiment, the `numpy` Python library was used, with a fixed random state seed set to 42.

[1]Available at https://drive.google.com/file/d/18bzjBNNeTWKGA38dm7xSOofRGQz9ZQ_D/view.
[2]The source code is available at https://github.com/haraleib/MMSR_GroupB_WS23

Table 1. Results of the Random Baseline retrieval system

Song	Artist
Beyond the Down	Black Label Society
Motion	Boy Harsher
Hole In My Soul	Kaiser Chiefs
Sans Logique	Mylène Farmer
Mirror	Kat Dahlia
Flash	Cigarettes After Sex
Long Cool Woman (In A Black Dress) - 1999 Remastered Version	The Hollies
Natural Harmony	The Byrds
Candy	Paolo Nutini
You Don't	GFOTY

Table 2. Results of the text-based TF-IDF retrieval system

Song	Artist	Similarity
Under My Umbrella	Margo Guryan	0.943
Teenage Love Affair	Alicia Keys	0.872
Blame It on the Boom Boom	Black Stone Cherry	0.849
Charlie Brown	Benito Di Paula	0.797
Walpurgisnacht	Faun	0.646
Barco a Venus	Mecano	0.639
Mariô	Criolo	0.612
Dirt	Alice in Chains	0.606
Shine Ya Light	Rita Ora	0.568
Auld Lang Syne (The New Year's Anthem)	Mariah Carey	0.536

2.2 Experiments

Out of three songs analyzed in total[2], the song *Waka Waka (This Time for Africa)* by *Shakira* served as the initial query for our experiments.

- **Random Baseline**: The random baseline system demonstrated purely random outputs without any noticeable pattern or relevance to the input queries as seen in Table 1.
- **Cosine similarity with TF-IDF**: Table 2 shows that the TF-IDF embeddings have a strong inclination towards matching lyrics and song titles. It suggests a dominance of literal textual similarity in the retrieval process, with genre relevance and thematic similarity playing a less prominent role.
- **Cosine similarity with BERT**: Table 3 demonstrates that the BERT embedding model predominantly focuses on genre similarity. This observation indicates that the BERT model, in this context, was more sensitive to genre-defining characteristics from the input data than to other aspects like thematic or literal textual similarity.
- **Cosine similarity with word2vec**: Table 4 implies that the word2vec embedding system's outputs were significantly influenced by thematic similarity at a conceptual level. This system seemed to capture the broader themes and ideas in the music tracks more effectively than the literal text.

Table 3. Results of the text-based BERT retrieval system

Song	**Artist**	**Similarity**
Sol da Liberdade	Daniela Mercury	0.649
El Vals del Obrero	Ska-P	0.629
Breakin'...There's No Stopping Us	Ollie & Jerry	0.618
BEAUTIFUL HANGOVER	Bigbang	0.618
Auf Anderen Wegen	Andreas Bourani	0.611
Feel Good Inc.	Gorillaz	0.608
VIVID	BROCKHAMPTON	0.606
The Bomb	Pigeon John	0.603
Mariô	Criolo	0.602
Free Me	Joss Stone	0.601

Table 4. Results of the text-based word2vec retrieval system

Song	**Artist**	**Similarity**
Blame It on the Boom Boom	Black Stone Cherry	0.861
Metaphors	San Cisco	0.847
Under My Umbrella	Margo Guryan	0.847
Baby's on Fire	Die Antwoord	0.843
Bamboreea	Inna	0.843
Royal	Waterparks	0.841
God Lives Through	A Tribe Called Quest	0.840
Kiss This	The Struts	0.839
Patterns	Simon & Garfunkel	0.839
Sorry - Latino Remix	Justin Bieber	0.838

It should be noted that the presented findings generalize across all of our experiments.

3 CONCLUSION

The choice of word embedding has a significant impact on the performance and relevance of content-based music retrieval systems. Each of the systems evaluated in our report exhibited characteristic behaviors and priorities in their outputs, which may or may not be relevant to the users inherently desired query.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805 cite arxiv:1810.04805.

[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK. http://nlp.stanford.edu/IR-book/information-retrieval-book.html

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781

[4] Marta Moscati, Emilia Parada-Cabaleiro, Yashar Deldjoo, Eva Zangerle, and Markus Schedl. 2022. Music4All-Onion – A Large-Scale Multi-Faceted Content-Centric Music Recommendation Dataset. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 4339–4343. https://doi.org/10.1145/3511808.3557656