Similarity Measures for Music Retrieval

Federico Newton

Supervisor: Luis A. Leiva
University of Luxembourg, Luxembourg

Abstract. We analyze six different similarity metrics for music retrieval on three public datasets. We use the same set of features to characterize songs and the same approach to compute ground-truth labels. We found that Cosine Similarity is often the best performer, followed by Chebyshev Distance. We also analyze the intersection of Precision and Recall curves to seek a balance between retrieval performance and length of recommendation lists. Our experiments can inform researchers interested in selecting the most appropriate metric for music recommender systems.

Keywords: Similarity · Dissimilarity · Recommender Systems

1 Introduction

Music recommender systems (RecSys) are widely used in popular streaming platforms such as Spotify or SoundCloud. Any music RecSys relies on some similarity measure or metric to determine how similar songs are. Therefore, the choice of this metric can have a strong impact on RecSys performance.

Surprisingly, there is no systematic analysis that compares various similarity measures for music RecSys. Previous work has focused on images [7, 2] and text [4, 6] or both [5], for example. In this paper, we focus on a range of features proposed by previous work that reflect song characteristics. We use the well-known Precision and Recall evaluation measures to effectively evaluate retrieval performance. We also analyze the most suitable length of recommended music lists, based on the intersection of Precision and Recall curves.

2 Materials and Method

We evaluate music retrieval performance according to six (dis)similarity measures [8]: Cosine Similarity, Euclidean Distance, Manhattan Distance, Canberra Distance, Chebyshev Distance, and Mahalanobis Distance. To promote generalizability, we chose three datasets for evaluation: cal500 [9], DEAM [1], and SiTunes [3]. These datasets contain user feedback and the same song features, based on Spotify's API, such as duration, loudness, BPM, etc.[1] For each dataset,

[1] All song features are listed and described at `https://developer.spotify.com/documentation/web-api/reference/get-audio-features`

user feedback was aggregated into an average Song-Likeness score for each song, which serve as ground-truth labels for evaluation.

We include feature selection as part of our evaluation pipeline, aiming at improving retrieval performance. Concretely, we applied the Benjamini-Hochberg procedure, which controls the False Discovery Rate (FDR) using an $\alpha = 0.05$.

The recommendation algorithm is k-nearest neighbors (k-NN). We use half of the data for querying the other half of the data. For each query song in a dataset, we compute a range of recommended songs, from one up to the maximum number of songs in the dataset. To account for class imbalance of positive and negative Song-Likeness scores, we use weighted Precision and Recall.

3 Results

Evaluation results are shown in Figure 1. The x-axis is normalized in the [0,100%] range, to ease cross-dataset comparison, and represents the percentage of recommendations retrieved (k). We can observe that in two out of the three datasets (DEAM and SiTunes) the Cosine Similarity reached the highest scores of both Precision and Recall, outperforming the other measures. This suggests that Cosine Similarity is the best choice of similarity measure when it comes to song recommendations. The Chebyshev Distance showed mixed results, as it performed best in the cal500 dataset, but achieved very poor results in the other two datasets. This indicates that this measure may not be reliable across datasets.

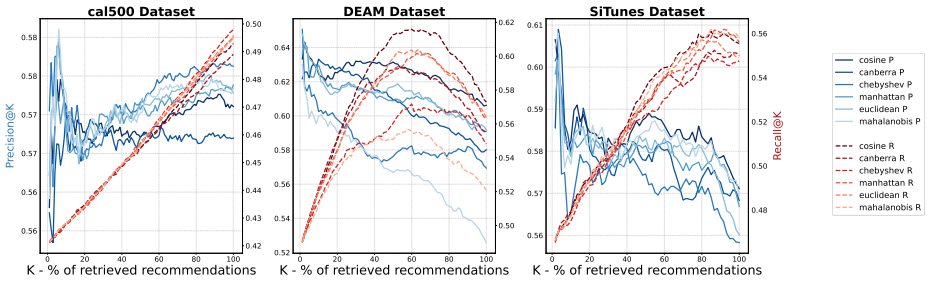

Fig. 1: Retrieval performance results. Legend: blue solid lines denote Precision, whereas red dashed lines denote Recall.

Furthermore, by analyzing the intersection of Precision and Recall curves, we can find a balance between both metrics. This balance, depending on the similarity measure, ranges often between 30–50% of the total amount of the dataset size. It indicates a trade-off for the music RecSys to generate accurate recommendations while also retrieving relevant songs. It also provides an informed choice to determine the maximum number of songs to present the user with, in a way that it is both accurate and not too overwhelming to scan.

References

1. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PLOS ONE **12**(3), 1–22 (03 2017). https://doi.org/10.1371/journal.pone.0173392
2. Flores-Pulido, L., Starostenko, O., Rodríguez-Gómez, G., Portilla-Flores, A., Mora-Lumbreras, M.A., Albores-Velasco, F.J., Sánchez, M.L., Cuamatzi, P.H.: Similarity metric behavior for image retrieval modeling in the context of spline radial basis function. In: Batyrshin, I., Sidorov, G. (eds.) Advances in Soft Computing. pp. 443–451. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25330-0_39
3. Grigorev, V., Li, J., Ma, W., He, Z., Zhang, M., Liu, Y., Yan, M., Zhang, J.: Situnes: A situational music recommendation dataset with physiological and psychological signals. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. p. 417–421. CHIIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3627508.3638343
4. Kumar, N., Yadav, S.K., Yadav, D.S.: Similarity measure approaches applied in text document clustering for information retrieval. In: 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). pp. 88–92 (2020). https://doi.org/10.1109/PDGC50313.2020.9315851
5. Malali, N., Keller, Y.: Learning to embed semantic similarity for joint image-text retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 10252–10260 (2022). https://doi.org/10.1109/TPAMI.2021.3132163
6. Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. In: Amati, G., Carpineto, C., Romano, G. (eds.) Advances in Information Retrieval. pp. 16–27. Springer Berlin Heidelberg, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71496-5_5
7. Patel, B., Yadav, k., Ghosh, D.: State-of-art: Similarity assessment for content based image retrieval system. In: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC). pp. 1–6 (2020). https://doi.org/10.1109/iSSSC50941.2020.9358899
8. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y.: A comparison study on similarity and dissimilarity measures in clustering continuous data. PLOS ONE **10**(12), 1–20 (12 2015). https://doi.org/10.1371/journal.pone.0144059
9. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech, and Language Processing **16**(2), 467–476 (2008). https://doi.org/10.1109/TASL.2007.913750