# KNOW THY NEIGHBOR: COMBINING AUDIO FEATURES AND SOCIAL TAGS FOR EFFECTIVE MUSIC SIMILARITY

*Alexandros Nanopoulos*

University of Hildesheim, Germany
nanopoulos@ismll.de

*Ioannis Karydis*

Ionian University, Corfu, Greece
karydis@ionio.gr

## ABSTRACT

Measuring similarity of two musical pieces is an ill-defined problem for which recent research on contextual information, assigned as free-form text (tags) in social networking services, has shown to be highly effective. Nevertheless, approaches based on contextual information require adequate amount of tags per musical datum in order to be effective. In the case of the so called "cold-start" problem, this assumption is not valid for several music data. In this paper, we address this problem by proposing a combination of the audio and the tag feature space of musical data. The application of the proposed combination for musical data lacking contextual information is shown, through experimental results with real musical data, to evaluate more accurately their similarity than the use of solely audio-based similarity.

***Index Terms***— audio similarity measurement, social tags, "cold-start" problem, contextual knowledge, audio-tag feature space combination

## 1. INTRODUCTION

Musical similarity has a central role in Music Information Retrieval (MIR) tasks, such as in commercial music dissemination and recommender systems, playlist generation, query by example, or finding common musical patterns.

The measurement of similarity between musical pieces is widely accepted to be hard to define in strictly objective terms [1]. Audio-based methods use features extracted from the musical signals, e.g., mel frequency cepstral coefficients. The performance of audio-based methods, however, has been identified to have reached a limit characterised as "glass ceiling" [2].

To overcome "glass ceiling", contextual knowledge is involved in the form of genre labeling, mood description, participation in playlist, as well *social tags*, i.e., free text labels assigned by humans [3]. Social tags constitute an almost unique source of human-generated information that cannot be attained by features extracted from audio. The information conveyed by social tags has been described to be of high importance to MIR [3, 4, 5], whereas recent research [6] has shown that measuring musical similarity based on tags is usually more accurate than audio-based methods.

Nevertheless, use of social tags comes with the prerequisite that these exist in an adequate amount, an assumption not always true. A common problem, referred to as "cold-start", pertains to newly released tracks or tracks of limited popularity that present a diminished number of tags assigned to them. According to Lamere [3], this problem is critical when tags are to be used for long-tail music discovery. Thus, in cases tags are few or non existent, the use of audio-based similarity measures is the only plausible alternative, despite the inferior accuracy of their computed similarity.

To address the aforementioned problem, we propose a novel approach denoted as *Know-Thy-Neighbor* (KTN). KTN exploits musical data for which adequate tags are available, for the purpose of combining the resulting audio feature and tag feature spaces. This combination is then applied for musical data that lack tags, thus, addressing the "cold-start" problem by providing more accurate similarity measurement and avoiding the "forced" use of audio-based similarity measures. We perform a thorough experimental evaluation with real data crawled from music web services (Last.fm and iTunes), the results of which indicate the clear benefits offered by the proposed method compared to the plain use of audio-based similarity measures for the case of "cold-start" problem.

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the formation of tag and audio feature spaces and details the proposed method, whereas Section 4 presents the experimental results. Finally, the paper is concluded in Section 5.

## 2. RELATED WORK

The superiority of methods that measure musical similarity based on tags is described by McFee et al. [6]. To address the lack of adequate amount of social tags, a number of research works focuses on combination of the tag feature space with other spaces. The work of Wang et al. [7] studies the problem of combining tags and audio content for artistic style clustering by proposing a language model that makes use of both data sources. In contrast to our work, Wang et al. focus on ameliorating artistic clustering performance by utilisation of both audio and tag spaces.

Research in *metric learning* has recently started to attract attention in the MIR domain. Metric learning can be utilised in order to exploit musical data for which adequate contextual information (tags) are available for the purpose of learning an effective mapping from the audio feature space to the tag feature space. This mapping can then be applied for musical data that have no or limited contextual information. McFee et al. [6] propose learning a content-based similarity for collaborative filtering. Their work focuses on optimising similarity for ranking, that is, similarity is evaluated according to the ranked list of results in response to a query example by use of the Metric Learning to Rank algorithm.

Researchers in [8, 9, 10] make use of mainly content-based audio analysis, among other methods, for the purposes of tagging a musical datum, also known as *autotagging*. These works are tackling partially the problem we focus in our study, since these could perform tag prediction for tracks without an adequate number of tags. We consider these approaches complementary to our proposed method, since our objective is not to predict tags for tracks, but to combine the tags available in amply tagged tracks with audio content of little or not at all tagged musical data for the purpose of computing more accurately their similarity. Finally, Kim et al. [11] perform autotagging using inter-track similarity based on sources such as user preference data, social tags, web documents, and audio content.

## 3. THE PROPOSED METHOD FOR COMBINING AUDIO AND TAG FEATURE SPACES

This section describes the approach we propose for combining audio features and tags. In Section 3.1, we first describe the formation of the tag and audio feature spaces, followed by Section 3.2 that provides a basic experimental comparison between them. Finally, Section 3.3 presents the proposed KTN method.

### 3.1. Tag and Audio Feature Spaces

We developed our experimental framework by accumulating the following three different types of data:

**Audio:** Audio data were harvested using the iTunes API[1]. Track selection was based on the cumulative highest popularity tags offered for the track in Last.fm[2] by selecting the 50 top rank tracks for each top rank tag. The data gathered contain $5,459$ discrete tracks and each track is a 30 second clip of the original audio, an audio length commonly considered in related research [7, 12].

**Social tags:** For each collected track, the most popular tags assigned to it were gathered using the Last.fm API, resulting in $84,334$ discrete tags, each track having on average 64 discrete tags assigned to it. Although Last.fm had a very large number of tags per track, our selection was based on the number of times a specific tag has been assigned to a track by different users. Thus on average the tags selected have been assigned 11 times by different users on a track.

**External metadata:** For each track, its respective metadata concerning the track's title, playing band, album title and genre were also collected. In contrast to audio and social tags, these external metadata where at *no* point used in the algorithms described herein. Their usage was merely as means for evaluating the accuracy of computed similarity. In the following, we focus on genre information, which is commonly used for evaluating similarity measures [12].

Based on the above data, we can develop two different feature spaces:

**Audio feature space:** We examined the commonly used Mel Frequency Cepstrum Coefficients (MFCCs). Following [13], the audio-based similarity in the created feature space is computed by first creating a cluster model and, then, using of a variation of the Earth Mover's Distance. All calculations on the audio content data was achieved using the MA toolbox [14].

**Tag feature space:** Tags are pre-processed to remove stop words and stemmed with the well-known Porter algorithm. Finally, Latent Semantic Analysis (LSA) is applied based on the Singular Value Decomposition (SVD) of the tag-track matrix, which produces a reduced dimensional representation that emphasises the strongest relationships and discards noise. In the sequel, we set as 20 the default value of dimensions for the SVD method. In the resulting space we compute similarity based on the cosine measure.

The comparison of all examined similarity measures is done with the calculated precision based on the $k$ nearest neighbors ($k$-NN) of query tracks: for each query track we measured the fraction of its resulting $k$-NNs that share the same genre with the query track. In the sequel, we set as default value $k = 10$. For each experiment we randomly select 80% of the data as training data and the remaining 20% act as testing data. Each experiment is repeated 30 times and the results are averaged.

### 3.2. Comparing Tag and Audio Feature Spaces

We first examine our assumption, that similarity in the tag feature space outperforms similarity in the audio feature space. It should be noted that only for this experiment, we assume that the "cold-start" problem does not exist, i.e., that the tags of the query tracks are *known*. The resulting precision w.r.t. the number, $k$, of queried nearest neighbors is displayed in Figure 1. As expected, similarity computed in the tag feature space clearly outperforms similarity in the audio feature space. This fact verifies the main motivation of our proposed method, which is described in the following.

[1] www.apple.com/itunes

[2] www.last.fm

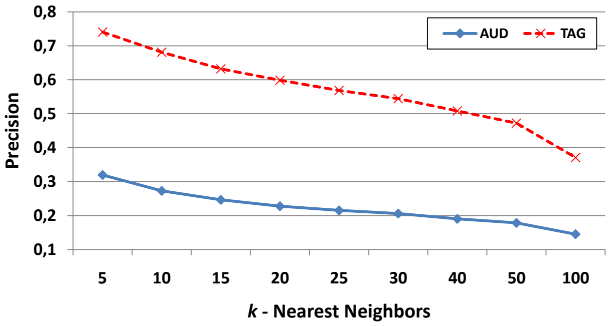

**Fig. 1**. Comparison of tag and audio feature spaces.

### 3.3. Proposed Method

Given is a collection, $D$, of tracks. For each query track, which does not belong in $D$, our task is to identify its $k$ nearest neighbors among the tracks of $D$ ($k$ is user-defined). To examine the "cold-star" problem, we henceforth assume that adequate tags exist for the tracks in the training collection $D$, however this does not hold true for the query tracks. Thus, our focus is on examining how effectively can the proposed method compute similarity for tracks that lack social tags.

The proposed method, denoted as Know-Thy-Neighbor (KTN), commences with a *training* phase in which the tracks of $D$ act as training data for which weights are learned. The weight of each such track is defined as its number of $n$-occurrences, i.e., the number of times it appears among the $n$ nearest neighbors of the rest tracks in $D$. The $n$-occurrences are initialised to zero and computed as follows. We find separately for each track in $D$ a list: (i) $L_1$ of its $n$ nearest neighbors based on the audio feature space, and (ii) $L_2$ of its $n$ nearest neighbors based on the tag feature space. We increase the number of $n$-occurrences for each track in $L_1 \cup L_2$, that is, we treat the two feature spaces uniformly. Regarding the value of $n$, based on empirical investigations we found that the proposed method is not sensitive to it and we retain $n = 10\%$ of the size of collection $D$ as default value for $n$. Query tracks are *not* involved during the weight learning.

For each query track, KTN searches initially in the audio feature space and finds its $k_1$ nearest neighbors, where $k1 \geq 1$ is a parameter of the method. From these $k_1$ neighbors, we select the single one with the highest weight. Next, KTN searches in the tag feature space in order to find the $k$ nearest neighbors of the selected track. These $k$ nearest neighbors are finally returned as the answer for the query track.

The intuition behind KTN is as follows: since not adequate tags exist for the query track, KTN identifies first in the audio feature space a representative neighbor for the query track. The representativeness of tracks in the audio feature space is determined by their learned weights, i.e., their $n$-occurrences, which promotes tracks that have the property of being popular nearest neighbors. This property is defined as *hubness* and has been recently analysed in [15]. The role of parameter $k_1$ is to maintain proximity information in the audio space (otherwise, tracks with global hubness would prevail). Since the selected neighboring track is considered as representative, its $k$ nearest neighbors based on the tag feature space are returned as an accurate estimation of the actual nearest neighbors of the query song. Therefore, KTN combines effectively both feature spaces to address the "cold-start" problem.

## 4. EXPERIMENTAL EVALUATION

We compare experimentally the proposed method, denoted as KTN, with the computation of similarity based solely on the audio feature space, denoted as AUD (i.e., AUD represents the "forced" use of audio-based similarity in the case of "cold-start"). We use the experimental framework described in Section 3.1.

Figure 2 presents the resulting precision w.r.t. $k$ nearest neighbors. For KTN two different $k_1$ values are examined: $k_1 = 1$ and $k_1 = 3$. It should be noted that $k_1 = 1$ leads to ignoring the learned weights (since the weights are used to compare the $k_1$ neighbors in the audio feature space). Evidently, KTN compares favorably to AUD. Moreover, the use of weights (when $k_1 = 3$) presents an advantage compared to the case these are ignored (when $k_1 = 1$). Using double t-tests, all differences in Figure 2 have been found statistically significant at level 0.05 for all examined $k$ values.

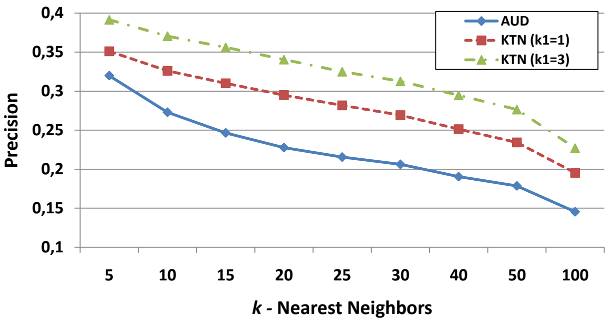

**Fig. 2**. Comparison between KTN and AUD.

Next, we examined the impact of the $k_1$ parameter used in KTN, where $k$ was set to 10. Figure 3 shows that a small $k_1 > 1$ value is better than setting $k_1 = 1$. Nevertheless, larger $k_1$ values reduce the performance of KTN, due to the locality in the audio space that is not any more preserved.

## 5. CONCLUSION

Motivated by the good performance of music similarity measures that are based on social tags, in this paper we proposed a novel approach for the combination of the feature space that is defined by social tags with the space defined by audio features. The proposed method is suitable in the case of the so

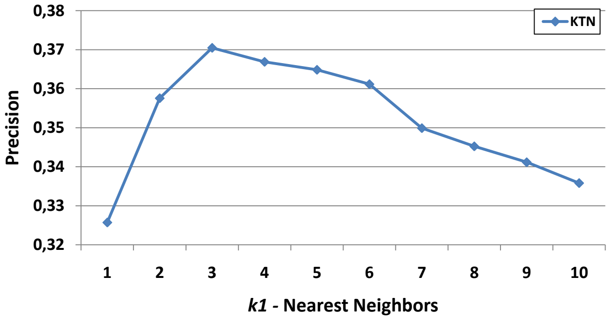

**Fig. 3**. Impact of $k_1$ parameter on KTN.

called "cold-start" problem that results from the lack of adequate tags. Our proposed method avoids the "forced" use of solely audio-based similarity measures when measuring music similarity and utilises available contextual knowledge in the form of social tags, which is known to be quite effective in MIR.

The proposed methodology is shown to be effective in comparison to the audio-based similarity computation w.r.t. precision of the resulting similarity measures. This is verified through experimental results with real data, which illustrate the suitability of the proposed method.

In future work, we plan to examine the addition of sources of contextual information, such as textual information from musical blogs, shared playlists, etc. as the content of these sources is equally of great importance to MIR.

## 6. REFERENCES

[1] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," in *Proc. International Society for Music Information Retrieval (ISMIR08) Conf.*, 2008, pp. 148–153.

[2] E. Pampalk, "Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns," in *in Proc. International Symposium on Music Information Retrieval (ISMIR06) Conf.*, 2006.

[3] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[4] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

[5] Alexandros Nanopoulos, Dimitrios Rafailidis, Panagiotis Symeonidis, and Yannis Manolopoulos, "Musicbox: Personalized music recommendation based on cubic analysis of social tags," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 2, pp. 407–412, 2010.

[6] B. McFee, L. Barrington, and G. Lanckriet, "Learning similarity from collaborative filters," in *Proc. International Society for Music Information Retrieval (ISMIR10) Conf.*, 2010.

[7] D. Wang, T. Li, and M. Ogihara, "Are tags better than audio? The effect of joint use of tags and audio content features for artistic style clustering," in *Proc. International Society for Music Information Retrieval (ISMIR10) Conf.*, 2010, pp. 57–62.

[8] D. Eck, T. Bertin-Mahieux, and P. Lamere, "Autotagging music using supervised machine learning," in *Proc. International Society for Music Information Retrieval (ISMIR07) Conf.*, 2007, pp. 367–368.

[9] M. Sordo, C. Lauier, and O. Celma, "Annotating music collections: How content-based similarity helps to propagate labels," in *Proc. International Society for Music Information Retrieval (ISMIR07) Conf.*, 2007.

[10] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transaction on Speech and Audio Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[11] J. H. Kim, B. Tomasik, and D. Turnbull, "Using artist similarity to propagate semantic information," in *Proc. International Society for Music Information Retrieval (ISMIR09) Conf.*, 2009, pp. 375–380.

[12] "Music information retrieval evaluation exchange," http://www.music-ir.org/mirex/wiki/MIREX_HOME.

[13] B. Logan and A. Salomon, "A music similarity function based on signal analysis," 2001.

[14] E. Pampalk, "A matlab toolbox to compute music similarity from audio," in *Proc. of 5th International Conference on Music Information Retrieval*, 2004.

[15] I. Karydis, M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Looking through the glass ceiling: A conceptual framework for the problems of spectral similarity," in *Proc. International Society for Music Information Retrieval (ISMIR10) Conf.*, 2010.