AUDIO-TO-TAG MAPPING: A NOVEL APPROACH FOR MUSIC SIMILARITY COMPUTATION

Ioannis Karydis

Ionian University, Corfu, Greece
karydis@ionio.gr

Alexandros Nanopoulos

University of Hildesheim, Germany
nanopoulos@ismll.de

ABSTRACT

Similarity measurement between musical pieces is a hard problem. Recent research on contextual information assigned (tags) in social networking services has shown to be highly effective in measuring musical similarity. Nevertheless, such an approach requires adequate amount of tags assigned to each musical datum. In the case of the so called "cold-start" problem, this assumption is not valid for several music data. Herein, we address this problem by proposing the utilisation of a learning mechanism that maps musical data from audio feature space to tag feature space. The developed mapping can be applied to musical data with no or limited contextual information, in order to more accurately evaluate similarity and avoid the sole use of audio-based similarity measures that may affect the similarity measurement quality. Experimental results with real musical data illustrate the substantial gains of the proposed method.

Index Terms— audio similarity measurement, social tags, "cold-start" problem, metric learning

1. INTRODUCTION

Measuring the similarity between two musical pieces is widely accepted to be a hard problem, as it is difficult to be defined in strictly objective terms [1]. For this reason, musical similarity is often based on subjective criteria, for which, however, contextual knowledge becomes an important factor [2]. It is possible to consider several sources of contextual information that can be utilised for musical data. Since human assigned social tags present several advantages in Music Information Retrieval (MIR) [3], we henceforth focus our examination on these.

Despite the inherent difficulties in assessing music similarity, its output is of great importance to numerous areas of MIR. Based on music-similarity measures, listeners are able to perform query by example tasks, musicologists identify common patterns between writers or writer development based on other writers, commercial music dissemination offer better suggestions on potential buyers using recommender systems, and music producers utilise special applications that produce template playlists for their work.

Some methods compute music similarity using objective metadata, e.g., composer name, song title, etc. However, such methods are in cases not as helpful as content-based methodologies are, since metadata require prior knowledge of data that is not conveyed by listening, may be potentially unavailable and have limited scope due to usage of pre-defined descriptors. Content-based similarity has been under extensive research [4, 5, 6, 7] focusing on features extracted from the audio content, which express different attributes of a musical datum. Nevertheless, it is becoming acceptable that the performance of content-based music similarity is reaching a limit that is characterised as "glass ceiling" [4].

The widespread penetration of "Web 2.0" and its increased social interactivity allowed the "wisdom of the crowds" to be available thought one of the most commonly attributed practice of users in social-networking websites, the assignment of tags that offer rich contextual information. The information conveyed by tags on musical data is of high importance to MIR [2, 8, 9] as it is an almost unique source of human-generated information that cannot be attained by audio-content processing.

Bearing in mind the aforementioned subjectivity of musical similarity and the nature of user assigned tags on musical data, it comes as no surprise that methods measuring musical similarity based on tags are frequently more accurate than content-based methods [3]. However, measuring musical similarity based on tags comes with the prerequisite that an adequate number of tags exist for the songs whose similarity is being measured. This assumption, however, is not always true, e.g., newly released songs or songs of limited popularity commonly present diminished number of tags assigned to them. This problem is also known as the "cold-start" problem that becomes critical when tags are to be used for long-tail music discovery [2].

1.1. Motivation

Although contextual information sources, such as social tags, present several benefits for measuring music similarity, the "cold-start" problem prevents their usage in several cases. When lacking contextual information, the use of audio-based similarity measures is a plausible alternative. Nevertheless, "forcing" the invocation of audio-based similarity can happen

at the cost of diminished performance in terms of the accuracy of the computed similarity.

What is, thus, required, is a novel approach that will incorporate the beneficial characteristics of the context domain (social tags) into the content domain (audio features). A principled learning approach can allow us to develop a *mapping mechanism* from the feature space that is defined by audio content to the feature space that is defined by the contextual information in the form of social tags. For this purpose, we aim 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. Such an approach can, thus, address the "cold-start" problem by providing more accurate similarity measurement and avoiding the "forced" use of audio-based similarity measures.

1.2. Contribution and Paper Organisation

In this paper we propose a novel method for improving the accuracy of similarity measurement for musical data that have not been assigned or present little contextual information ("cold-start" problem). Our contributions are summarised as follows:

Initially, we propose a novel method, called *Audio-to-Tag mapping* (A2T), that takes advantage of existing contextual information in the form of social tags, and learns a mapping between the audio and tag feature spaces. Our approach formulates the mapping problem as learning a matrix that minimises the relative entropy between the two kernel matrices which compute the similarity between the two aforementioned feature spaces. In addition, we describe the effective application of A2T for the case of the "cold-start" problem, for which A2T can use audio features to map songs to the tag feature space and perform similarity measurement within it, thus avoiding to rely solely on audio-based features. Finally, we perform experimental evaluation with real data crawled from music web services (Last.fm and iTunes), indicating the clear benefits offered by A2T compared to the plain use of audio-based similarity measures for the case of "cold-start" problem.

It should be noted that the proposed methodology is not similar to the automated tag creation for musical pieces, also known as "auto-tagging". Our proposed method does not produce a list of tags, on which similarity would then be measured, for a piece with no or limited tags based on similarity of other spaces.

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 details on content and context domains and provides a formal definition of the problem investigated in our work, whereas Section 4 presents the proposed method for mapping from audio to context. Section 5 presents and discusses the experimentation and results obtained, while the paper is concluded in Section 6.

2. RELATED WORK

Research utilising metric learning has recently started to attract attention in the MIR domain. Slaney et al. [1] describe means to embed acoustic features of musical data into a metric space. Aiming at the pairwise Euclidean distance between two songs to reflect semantic dissimilarity, their approach allows distance-based analysis, such as *k*-NN classification, to detect similar songs within a collection. Differently from our proposal, [1] assumes similarity based on the original content space using the Mahalanobis distance while no use of context knowledge, in the form of social tags, is applied therein. Metric learning has also been utilised by McFee et al. [3] in order to learn 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. In comparison to our work, in [3] the main approach is applied to three different sources of information without having any mapping between sources, and the proposed algorithm is constrained for the case of optimising similarity for the purposes of ranking.

In addition, recently, research in combining context with content data has started being explored. The work of Wang et al. [10] studies the problem of combining tags and audio contents for artistic style clustering by proposing a language model that makes use of both data sources. In contrast to our work, Wang et al. do not utilise metric learning.

Researchers in [11, 12, 13] make use of mainly content-based audio analysis, among other methods, for the purposes of tagging a musical datum, also known as auto-tagging. These works are tackling partially the problem we focus in our study, since they could perform tag prediction for songs without an adequate number of tags. We consider these approaches complementary to our proposed method, since our objective is not to predict tags for songs, but to map songs effectively from the audio feature space to the tag feature space for the purpose of computing more accurately their similarity.

3. PROBLEM DEFINITION: MAPPING FROM AUDIO FEATURE SPACE TO TAG FEATURE SPACE

3.1. Similarity based on Audio Features

The audio feature space usually comprises of features extracted from audio waveforms. We follow the commonly used assumption that a collection of songs is represented within the audio feature space in a matrix form, where each column represents a song and rows are the features extracted, i.e., each song is represented as a column vector. Accordingly, let $X = (x_1, x_2, \ldots, x_n)$ denote a collection of n input songs, each x_i being a column vector of m dimensions containing the extracted audio features and thus X be a matrix of size $m \times n$.

Based on the matrix representation of the audio feature space, similarity between songs can be defined in several

ways, e.g., using common measures like the Euclidean distance or cosine similarity. Our empirical investigations did not indicate significant differences between such common measures, partially due to their close relationship [14]. Therefore, in the sequel we use Equation 1 to define a kernel matrix S_A of all pair-wise similarities within the audio feature space, which computes the inner-product between each pair of songs.

$$S_A = X^T X \tag{1}$$

Using a matrix representation of items (songs in our case) in a feature space, and a similarity measure that is defined in this space (such as Equation 1), constitutes a fundamental approach in Information Retrieval and is also being commonly applied in MIR [7, 15]. It should be noted however, that such an approach does not generalise directly to similarity measures that are based on bag-of-frames representations and MFCCs. Based on our empirical investigations and on analogous results reported in the past [2, 8], a wide range of pure audio-based similarity measures (including bag-of-frame approaches) are in cases outperformed by similarity measures that are based on contextual information in the form of social tags. For this reason, and to simplify the presentation, we henceforth focus on the aforementioned matrix representation in the audio feature space and address the extension to bag-of-frames approaches as the main topic of our future work.

3.2. Similarity based on Tag Features

Following the matrix representation in the case of audio feature space, provisions need be made for the contextual information, that is, the tags assigned to each song. We initially consider the total number of tags, C, assigned to n given songs. We define a matrix, Y, where each row corresponds to a distinct tag and each column to a song. Accordingly, let $Y = (y_1, y_2, \ldots, y_n)$ denote tag vectors that represent the tags assigned to each song $y_i, 1 \leq i \leq n$. Each $y_i = (y_i^1, y_i^2, \ldots, y_i^C)$ is a vector with C elements.

Based on the representation of contextual information (in the form of social tags) with the Y matrix, all pairwise similarities can be computed with the kernel matrix S_T using Equation 2.

$$S_T = Y^T Y \tag{2}$$

As described also in Section 3.1, Equation 2 determines the similarity between each pair of songs based on their inner-product in the representation with the Y matrix. Although several other measures could be used, such as the cosine similarity, our empirical investigation did not indicate significant differences compared to the use of Equation 2. Thus, using Equations 1 and Equation 2 we have an analogous way to compute the similarity in the two different feature spaces, a fact that helps in the following to present the problem definition and its solution in a concise way.

Regarding the preprocessing of the contextual information, since tags are free-form text assigned by users, we have employed Latent Semantic Analysis (LSA) [16], in order to alleviate the problem of finding relevant musical data from search tags [2]. The fundamental difficulty arises when tags are compared to find relevant songs, as the task eventually requires the comparisons of the meanings or concepts behind the tags. LSA attempts to solve this problem by mapping both tags and songs into a "concept" space and doing the comparison in this space. For this purpose, we used Singular Value Decomposition (SVD) in order to produce a reduced dimensional representation of term-document matrix that emphasises the strongest relationships and reduces noise.

3.3. Problem Definition

Since computing similarity in the tag feature space is widely considered as advantageous compared to basing similarity solely on audio-features, in this study we address the following problem:

Problem Definition. We are provided with a collection of n songs for which we have their representations both in the audio feature space and in the tag feature space (i.e., we assume adequate contextual information for all n songs). For a given new song, we want to assess its similarity to the songs in the collection. We focus on the "cold-start" case and assume that not adequate contextual information, in the form of social tags, is available for the new song. Thus, we cannot compute its similarity to other songs within the tag feature space using directly the kernel of Equation 2. However, for the new song, we can determine the vector x that represents this new song within the audio feature space. Moreover, we can define a matrix, M, that maps vector x to Mx. The entries of M can be *learned* from the collection of n songs. In particular, following Equation 1, we define a derived kernel, S_D, which computes similarity after mapping all n songs from the audio feature space as follows:

$$S_D = (MX)^T(MX) = X^T M^2 X = X^T A X, \tag{3}$$

where we denote $M = A^{1/2}$. The learning task is defined in terms of minimising the distance between the kernel matrices S_D (Equation 3) and S_T (Equation 2). We measure the distance between the two kernel matrices based on the relative entropy $d(S_D||S_T)$ [17].

4. A2T: PERFORMING THE MAPPING FROM AUDIO FEATURE SPACE TO TAG FEATURE SPACE

As described in Section 3.3, our objective is to learn a mapping matrix, M, in order to minimise the distance $d(S_D||S_T)$ between the derived kernel (Equation 5) and the kernel of the context domain (Equation 2). Tsuda et al. [18] describe the computation of a distance between two positive definite matrices, based on the relative entropy. To use this result, and avoid the problem that S_T is singular[1], we use a common "smooth-

[1] Singularity of S_T results from the fact that, usually, $C < n$.

ing" technique to modify Equation 2 and compute based on Equation 4 a smoothed kernel, S_T^s, for the context domain:

$$S_T^s = Y^T Y + \lambda I_n, \quad (4)$$

where I_n is the identity matrix and n is the number of songs in the collection. The smoothing parameter $\lambda > 0$ allows a controlled alteration of Equation 2 in order to balance the trade-off between modifying the computed similarities in the tag feature space and avoiding singularity of S_T. Based on our empirical experience $10^{-2} \leq \lambda \leq 10^{-1}$ values are effective enough in order to avoid singularity and apply minimum data alteration. This characteristic of the λ parameter will be examined experimentally in Section 5.2.

Next, we can apply the result of Tsuda et al. [18] for S_D and S_T^s, as follows:

$$d(S_D||S_T^s) = \frac{1}{2}\left(tr((S_T^s)^{-1}S_D) + \log|S_T^s| - \log|S_D| - n\right), \quad (5)$$

where tr and $|\cdot|$ denote the trace and the determinant of a matrix, respectively.

Since $S_D = X^T A X$, we can cast the problem of computing the optimal A^* (please recall that $M = A^{1/2}$) as a minimisation problem in the following way:

$$A^* = \arg\min_{A \succeq 0} d(S_D||S_T^s) \quad (6)$$

The solution A^* is computed by setting the derivative of $d(S_D||S_T^s)$ (Equation 5) w.r.t. A equal to 0 and solving for A. Accordingly, we have:

$$A^* = \left(X(S_T^s)^{-1}X^T\right)^{-1} \quad (7)$$

Having computed A^*, for a new song that is represented with vector x in the audio feature space, we can compute its similarity to all other songs in the collection based on the derived kernel $S_D = X^T A^* X$. This way, S_D performs an effective mapping of all songs (including x) from the audio feature space (representation with matrix X) to the tag feature space (representation with matrix Y).

5. PERFORMANCE EVALUATION

5.1. Experimental Setup

For the purposes of performance evaluation of the proposed method we accumulated the data from iTunes and Last.fm web services.

Audio: Content data were harvested from iTunes using the iTunes API. Track selection was based on the cumulative highest popularity tags offered for the track in Last.fm by selecting the fifty 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 [10].

Social tags: For each track accumulated, the most popular tags assigned to these at the Last.fm were gathered using the Last.fm API. The data gathered contain $84,334$ discrete tags. Each track has 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 gathered from iTunes, its respective metadata concerning the track's title, playing band, album and genre were also stored. In contrast to the former two types of data, audio and social tags, the external metadata where at no point used in the algorithms proposed 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 in MIREX participations.

As far as the audio content data are concerned, the following content features were extracted: spectral centroid, spectral roll-off point, spectral flux, compactness, spectral variability, root mean square, fraction of low energy windows, zero crossings, strongest beat, beat sum, strength of strongest beat, thirteen first MFCC coefficients, ten first LPC coefficients and five first method of moments coefficients. Extraction was achieved using the *jAudio* [19] application for each entire musical datum producing thus a single content feature point of 39 dimensions per track.

For the social tags, each tag has been pre-processed in order to remove stop words that offer diminished specificity and additionally stemmed in order to reduce inflected or derived words to their stem using the algorithm described by Porter [20]. Moreover, tags were further processed using the LSA method as already described in Section 3.2 in order to minimise the problem of finding relevant musical data from search tags. To this end, the SVD method has been used in order to produce a reduced dimensional representation of term-document matrix that emphasises the strongest relationships and discards noise. Unless otherwise stated, the default value of dimensions for the SVD method was set to 50 dimensions.

For the evaluation of the similarity measure resulting by A2T, we use the precision resulting from the k nearest neighbors (k-NN) of a query song, i.e., for each query song we measure the fraction of its k-NN that share the same genre with the query song. In the sequel, we set as default value $k = 10$. For each experiment we randomly select 80% of the data as training data, i.e., songs that act as the collection. The remaining 20% act as testing data, i.e., query songs for which no social tagging information is assumed to exist ("cold-start" problem). Each experiment is repeated 30 times and the results are averaged.

5.2. Experimental Results

5.2.1. Tag-based vs. Audio-based

In the first experiment we examine our assumption that similarity in the tag feature space outperforms similarity in the audio feature space. Please notice that only for this experi-

ment, we assume that the "cold-start" problem does not exist, i.e., that the tags of the query songs are *known*. For this reason, no mapping is needed to be performed, since the similarity of query songs to the songs in the collection is computed using directly Equation 2. The purpose of this experiment is to verify the aforementioned assumption, which serves as the main motivating factor for developing the proposed A2T method.

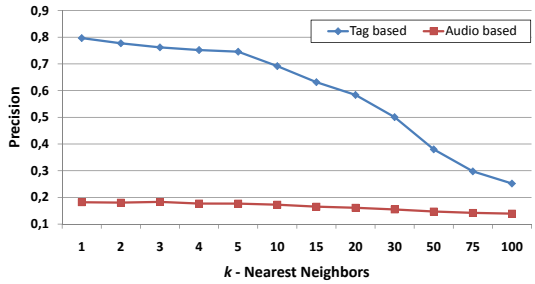

Fig. 1. Audio-based vs. tag-based similarity.

As seen in Figure 1 the Tag-based approach compares favorably to the Audio-based approach. This comes as no surprise, since the contextual information provided by tags is known to be very important for the purposes of MIR, as already discussed in Section 1, while the audio-based approach does not utilise any information from the context domain solely relying on the extracted audio features. Therefore, the motivation of A2T is to maintain the good characteristics of the tag feature space for query songs whose tags are *not* known.

5.2.2. *A2T vs. Audio-based*

In the next experiment, we examine the ability of A2T to learn an effective mapping and attain improved accuracy in computed similarity compared to the similarity that is based solely on audio features. Figure 2 presents the resulting precision for varying k number of nearest neighbors (for A2T, λ parameter is set to 10^{-2}). Clearly, A2T presents the best performance in all cases. To verify the importance of the improvement resulting from A2T, we applied double t-tests and found the differences significant at level 0.05 for all examined k values.

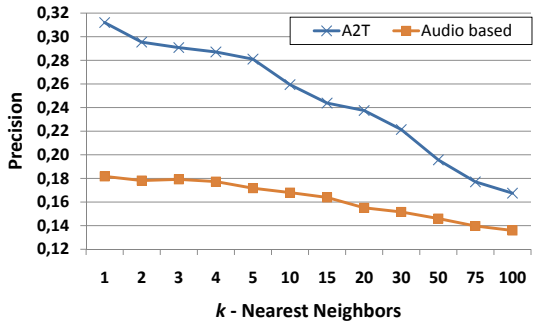

Fig. 2. Audio-based similarity vs. A2T Mapping.

Next, we aim in providing further insight as to the reasons why A2T can attain similarity with superior accuracy. We examine the ability of the similarity measure resulting by A2T to find k-NN songs that are close to the k-NN songs that would be found, if we could use directly the representation of query songs in the tag feature space (i.e., using Equation 2, resulting in the method called Tag-based in Section 5.2.1, which assumes that such representation is possible for the query songs and comprises an "upper bound" for the mapping procedure). For each query song, we determined its k-NN songs based on the similarity measure that results from A2T. Then, we measure the collective (summation) similarity of the k-NN songs and the query song directly in the tag feature space (i.e., assuming the availability of the corresponding representation for the query song and using Equation 2). The same measurement was performed for the case where the k-NN songs of the query song where determined using the audio-based approach. Figure 3 presents the percentage of difference in these collective similarities between A2T and audio-based approach (for A2T we examine 3 different cases for dimensions kept by SVD). The superiority of the A2T is clear, since the measured collective similarity for A2T is always better (up to 100% for small k values) compared to that of the audio-based approach. This indicates the effective mapping that A2T manages to perform, because the query songs have k-NN songs that are closer to those that would have been computed by the "upper-bound" method, which assumes full knowledge of the representation of query songs in the tag feature space.

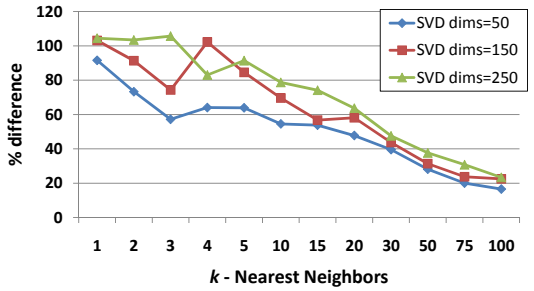

Fig. 3. A2T vs. audio-based methods' resulting k-NNs % difference with tag-based approach's k-NNs.

5.2.3. *Sensitivity of A2T*

Finally, we examined the impact of the parameters used in A2T, namely the number of dimensions kept by the SVD algorithm, as described in Section 3.2, and the smoothing parameter λ with respect to the precision achieved by the resulting similarity measure. Figure 4 shows that an increase in the dimensions utilised in SVD has a clear augmenting impact on the precision of the resulting similarity. Still, for larger increase, the ability of SVD to emphasise the strongest relationships and discard noise in data, diminishes and so does the precision of the resulting similarity. In addition, the effect of the smoothing parameter is quite evident. λ values between 10^{-2} and 10^{-1} are offering the best trade-off point between data alteration and avoiding singularity, while higher values

show a decrease in precision, due to the reduced effect of the smoothing. In the best case the effect of λ offers an increase of approximately 5.8% in precision.

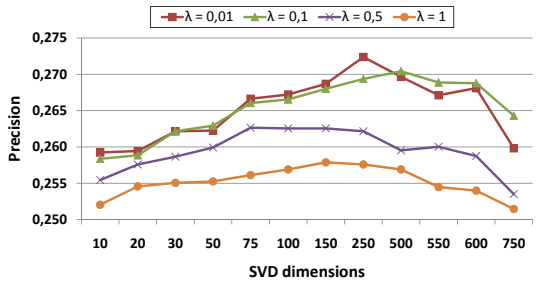

Fig. 4. Sensitivity of A2T to: i) dimensions kept by SVD and ii) λ.

6. CONCLUSION

In this paper, we proposed a novel approach for incorporating characteristics of the context domain of musical data, in the form of social tags, into the space defined by audio features, for the purposes of addressing the "cold-start" problem, differently than auto-tagging approaches. Our proposal avoids the sole usage audio-based similarity measures when measuring music similarity and utilises available contextual knowledge which is known to be quite effective in MIR.

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

In future work, we plan to examine audio similarity measures that are based on bag-of-frames approaches. For this reason, we will focus directly on matrices that represent the pair-wise similarity measures, i.e., not assuming representations as vectors within feature spaces.

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

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

[4] 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.

[5] B. Logan, D. P. W. Ellis, and A. Berenzweig, "Toward evaluation techniques for music similarity," *in Proc. International Conference on Multimedia & Expo*, vol. 2003, 2003.

[6] J. J. Aucouturier and F. Pachet, "Music similarity measures: Whats the use?," in *Proc. International Symposium on Music Information Retrieval (ISMIR03) Conf.*, 2003, pp. 157–1638.

[7] K. West and P. Lamere, "A model-based approach to constructing music similarity functions," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–10, 2007.

[8] 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.

[9] 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.

[10] 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.

[11] G. Tzanetakis and P. R. Cook, "Musical genre classification of audio signals," *IEEE Transaction on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[12] 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.

[13] 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.

[14] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between euclidean and cosine angle distance for nearest neighbor queries," in *Proc. ACM symposium on Applied computing*, 2004, pp. 1232–1237.

[15] D. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence, "The quest for ground truth in musical artist similarity," in *Proc. International Symposium on Music Information Retrieval (ISMIR02) Conf.*, 2002.

[16] S. T. Dumais, G. W. Furnas, T. K. Landauer, and S. Deerwester, "Using latent semantic analysis to improve information retrieval," in *Proc. Conference on Human Factors in Computing*, 1988, pp. 281–285.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[18] K. Tsuda, S. Akaho, and K. Asai, "The em algorithm for kernel matrix completion with auxiliary data," *Journal of Machine Learning Research*, vol. 4, pp. 67–81, 2003.

[19] D. McEnnis, C. McKay, and I. Fujinaga, "jaudio: Additions and improvements," in *Proc. International Society for Music Information Retrieval (ISMIR06) Conf.*, 2006, pp. 385–6.

[20] M. F. Porter, "The porter stemming algorithm," .