Comparative analysis of content-based and context-based similarity on musical data

C. Boletsis, A. Gratsani, D. Chasanidou, I. Karydis, and K. Kermanidis

Dept. of Informatics, Ionian University, Kerkyra 49100, Greece
{c10bole,c10grat,c10chas,karydis,kerman}@ionio.gr

Abstract. Similarity measurement between two musical pieces is a hard problem. Humans perceive such similarity by employing a large amount of contextually semantic information. Commonly used content-based methodologies rely on information that includes little or no semantic information, and thus are reaching a performance "upper bound". Recent research pertaining to contextual information assigned as free-form text (tags) in social networking services has indicated tags to be highly effective in improving the accuracy of music similarity. In this paper, we perform a large scale (20k real music data) similarity measurement using mainstream content and context methodologies. In addition, we test the accuracy of the examined methodologies against not only objective metadata but real-life user listening data as well. Experimental results illustrate the conditionally substantial gains of the context-based methodologies and a not so close match these methods with the real user listening data similarity.

1 Introduction

For a classic rock lover, Led Zeppelin's "Kashmir" and Deep Purple's "Perfect Strangers", may be two similar songs while for a hip-hop lover the very same songs may be completely different and an association of Led Zeppelin's "Kashmir" with Puff Daddy's "Come with me" is quite possible. The aforementioned example portrays just one scenario of the purely subjective nature of music similarity assessment and the problem that its measurement poses [27, 9].

Despite the inherent difficulties in assessing musical similarity, its function is of high value to numerous areas of Music Information Retrieval (MIR) [9]. Based on music-similarity measures [9]: (a) listeners can query using performed or hummed parts, (b) music researchers can identify recurring parts in different works, (c) the music industry offers music discovery tools in order to support potential buyers, and (d) music professionals and amateurs can organise their music effectively.

Musical similarity depends on the characterising attributes of the musical data to be compared and thus has been focused on three key directions: the objective metadata accompanying the musical works, the actual musical content and the contextual information humans assign on everything music.

Objective metadata, such as the song title, the singer name, the composer name or the genre of a musical piece can be used to assess music similarity. However, methods using metadata are in some cases not effective since metadata may be unavailable, their use requires knowledge that is, in general, not conveyed by

listening, and in addition have limited scope, as these rely on predefined descriptors [9].

Content-based similarity focuses on features extracted from the audio content. This task appears as a common process for humans due to the powerful ability of the brain to utilise an enormous amount of contextually semantic information for the process of identifying similarities and differences between sounds as well as classifying these sounds [6, 21]. On the contrary, in automated computer systems, the equivalent process based on content extracted features is much more difficult as the attributes expressed by the extracted features are of very little or lacking any semantic meaning [21].

Contextual knowledge, on the other hand, is derived from the information that humans apply to music through the practice of appointment free-form text (a.k.a. tags) on musical data on the web. Based on the previously mentioned ability of the human brain to utilise contextual information for music similarity and the rich contextually semantic nature of the human-generated information that is assigned to the musical works, the important role of tagging in MIR comes as no surprise. Consequently, measurements of musical similarity based on tags are in cases [19, 9] reported more accurate than content-based measurements. However, contextual information is no panacea, as far as music similarity is concerned and a number of issues are reported [12] to burden its use.

1.1 Contribution & Paper Organisation

In this paper, we compare and evaluate content-based versus context-based approaches for measuring music similarity. The contribution of this work is summarised as follows:

- Perform large scale (20k tracks) similarity measurement using mainstream content and context methodologies.
- Measure the accuracy of the examined methodologies against not only metadata but real-life user listening data.

The rest of the paper is organised as follows. Section 2 describes background and related work, Section 3 provides a complete account of the similarity measurement methods examined. Next, Sectiom 4 describes the context-based similarity approach examined herein. Subsequently, Section 5 presents and discusses the experimentation and results obtained, while the paper is concluded in Section 6.

2 Related Work

Music information retrieval has been under extensive research in the last decade and *similarity measurement* has been at the very core of the research [16, 22, 23, 27, 3, 2, 5] due to its importance to numerous areas of MIR.

Content-based similarity has been the corner-stone of automated similarity measurement method in MIR and most research [24, 16, 22, 3, 2, 5, 11] is focused in this direction. Content-based approaches assume that documents are described

by features extracted directly from the content of musical documents. Accordingly, the selection of appropriate features is very important as meaningful features offer effective representation of the objects and thus accurate similarity measurements. The work of Pampalk [22, 25] on Single Gaussian Combined, as submitted to the MIREX 2006 [20] is of high importance as it achieved the highest score and in addition, in current literature, spectral measures are receiving an ever growing interest as these describe aspects related to timbre and model the "global sound". In the direction of content-based feature usage and in order to alleviate the burden of programming for the extraction of features, McEnnis et al. [17, 18] developed a feature extraction library.

In contrast to content-based attributes of the musical data, context-based information refers to semantic metadata appointed by humans. Initial research in this direction focused in mining information from the web [4, 10] for the purposes of artist classification and recommendation. Nevertheless, the widespread penetration of "Web 2.0" enabled web users to change their previous role of music consumers to contributors [8] by simply assigning tags information on musical data. The increased appeal of the tagging process led to the assignment of large amounts of such information on everything musical. Accordingly, research [12, 14, 15] expanded in this direction in order to measure the similarity of musical content. Lamere [12] explores the use of tags in MIR as well as issues and possible future research directions for tags. Finally, Levy and Sandler [14] present a number of information retrieval models for music collections based on social tags.

3 Content-based similarity

Content-based approaches assume that documents are described by features extracted directly from the content of musical documents [11]. In our analysis, we experiment with two widely known cases: (a) content feature extraction based on the jAudio application [17] that produces a set of, generic for the purposes of MIR, features and (b) the more MIR specific Single Gaussian Combined method, as implemented in the MA Toolbox Matlab library [22], that was shown to perform more than adequately in the MIREX contests.

3.1 Generic features

MIR processes depend heavily on the quality of the extracted audio features [18]. The performance of a classifier or other interpretive tool is defined by the quality of the extracted features. Thus, poor-quality features will result in the poor performance of the classifier. The extracted features can be portrayed as a "key" to the latent information of the original data source [18]. Since, in our study, we focus on the interpretive layer, we created and maintained a large array of features. For the extraction of these features the jAudio application was used.

jAudio is an application designed to extract features for use in a variety of MIR tasks [18]. It eliminates the need for reimplementing existing feature extraction algorithms and provides a framework that facilitates the development and deployment of new features [18].

jAudio is able to extract numerous basic features [17]. These features may be one-dimensional (e.g., RMS), or may consist of multi-dimensional vectors (e.g., MFCC's) [18]. Metafeatures are feature templates that automatically produce new features from existing features [18]. These new features function just like normal features-producing output on a per-window basis [18]. Metafeatures can also be chained together. jAudio provides three basic metafeature classes (Mean, Standard Deviation, and Derivative).

For the purposes of our experimentation we retained the following features: 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, first thirteen MFCC coefficients, first ten LPC coefficients and first five method of moments coefficients.

3.2 Targeted features

In order to proceed to the extraction of targeted features, we utilised the feature extraction process based on the Single Gaussian Combined (G1C) [23]. Initially, for each piece of music the Mel Frequency Cepstrum Coefficients (MFCCs) are computed, the distribution of which is summarised using a single Gaussian (G1) with full covariance matrix [22]. The distance between two Gaussians is computed using a symmetric version of the Kullback-Leibler divergence. Then, the fluctuation patterns (FPs) of each song are calculated [22]. The FPs describe the modulation of the loudness amplitudes per frequency bands, while to some extent it can describe periodic beats. All FPs computed for each window are combined by computing the median of all patterns. Accordingly, two features are extracted from the FP of each song, the gravity (FP.G) which is the centre of gravity of the FP along the modulation frequency dimension and the bass (FP.B) which is computed as the fluctuation strength of the lower frequency bands at higher modulation frequencies [22]. For the four distance values (G1, FP, FP.B and FP.G) the overall similarity of two pieces is computed as a weighted linear combination (normalised in [0,1]) as described in detail in [23].

4 Context-based similarity

As far as contextual information is concerned, as tags are free-form text assigned by users, it requires preprocessing. Accordingly we employed Latent Semantic Analysis (LSA) [7], in order to alleviate the problem of finding relevant musical data from search tags [12]. 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 the term-document matrix that emphasises the strongest relationships and reduces noise.

5 Performance Evaluation

In this section we experimentally compare the accuracy of the content and context based methods using as groundtruth both the metadata of the tracks and the

similarity provided Last.fm [13] web service based on real-life user listening data. We initially describe the experimental set-up, then present the results and finally provide a short discussion.

5.1 Experimental Setup

For the purposes of performance evaluation of the alternative methods to compute similarity we accumulated two datasets from web services. The first dataset, henceforth titled *dataset A*, comprises of data selected for their high volume of contextual information, tags, as assigned in the Last.fm. The aforementioned web service does in addition provide, for most of the tracks, other tracks that are similar to them, based on user listening data. Thus, the second dataset, henceforth titled *dataset B*, comprises of tracks that are similar to the tracks of dataset A, following the information provided by Last.fm.

- **Audio:** Content data were harvested from iTunes [1] using the iTunes API. Track selection for dataset A was based on the cumulative highest popularity tags offered for a track in Last.fm by selecting the fifty top rank tracks for each top rank tag. Track selection for dataset B was based on their similarity to the tracks of dataset A following the information provided by Last.fm. The data gathered contain $5,460$ discrete tracks for dataset A and $14,667$ discrete tracks for dataset B, retaining only the first 10 most similar tracks for each track of dataset A. Each track is a 30 second clip of the original audio, an audio length commonly considered in related research [28, 20].
- **Social tags:** For each track accumulated, the most popular tags assigned to it at Last.fm were gathered using the Last.fm API. The data gathered contain more than $165,000$ discrete tags. 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.
- **External metadata:** For each track gathered from iTunes, its respective metadata concerning the track's title, artist, album and genre were also stored. In contrast to the former two types of data, audio and social tags, the external metadata where merely used as a means for evaluating the accuracy of computed similarity. In following experimentation we focus on genre information, which is commonly used for evaluating similarity measures [20, 9].

As far as the audio content data is concerned, the representation of tracks in our experimentation is based on the following two schemes: (a) Content features: 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, first thirteen MFCC coefficients, first ten LPC coefficients and first five method of moments coefficients, as described in Section 3.1. Extraction was achieved using the *jAudio* [18] application for each entire musical datum producing thus a single content feature point of 39 dimensions per track. (b) Content features: Single Gaussian Combined (G1C) as described in Section 3.2. Extraction was achieved through *MA Toolbox*, a collection of Matlab functions that implement G1C, as described in [23]. Throughout the remainder of this paper, the latter scheme is used except when explicitly stated otherwise.

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 [26]. Moreover, tags were further processed using the LSA method as already described in Section 4 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 the 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.

Initially we tested the methodologies examined herein solely in dataset A. Accordingly, Figures 1 and 2 report results on similarity measurement accuracy just for dataset A. On the other hand, Figures 3, 4 and 5 present results concerning the incorporation of dataset B into the similarity measurement process, following the similarity results of Last.fm, in order to use it as a groundtruth. Thus, the intuitive result of using real user listening data as a groundtruth similarity is to observe the capability of the examined methodologies to measure similarity similarly to the manner real-life users would.

For the evaluation of the similarity between tracks, we used the precision resulting from the k nearest neighbors (k-NN) of a query song, i.e., for each query song we measured the fraction of its k-NN that share the same genre with the query song. In the cases that employ both datasets A & B, queries are selected from dataset A while similar matches are retrieved from both datasets.

5.2 Experimental Results

In the first experiment, Figure 1(left), we tested the accuracy of similarity measurement using solely the content of tracks from subset A. For this experiment we utilised the features extracted using the jAudio application representing thus each track by a 39 dimension vector. This experiment verifies that for a generic set of features, extracted from the content of a track, the mean precision is very low, serving thus as a key motivation factor for the development of methodologies that perform better. In the next experiment, we examined the attained accuracy in computed similarity utilising the features included in the MA-Toolbox. Figure 1 (right) presents the resulting precision for varying k number of nearest neighbors using the G1C features. As in the previous result, the initial setting accuracy provided by the MA-Toolbox is comparable to the accuracy provided by the generic set of features.

Continuing further, the next experiment presents the accuracy of the similarity measurement using the contextual information of the dataset A tracks. Figure 2 clearly shows that the accuracy of similarity measurement in the tag feature space outperforms similarity in the audio feature space. In addition, the effect of the SVD dimensionality reduction can also be seen: 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.

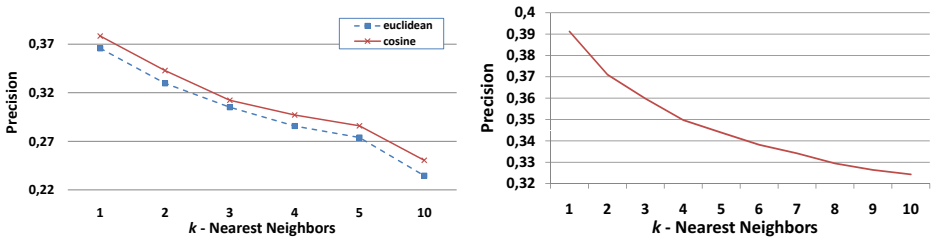

Fig. 1. Dataset A - content, mean precision vs. kNNs, using features extracted from jAudio (left) and from MA-Toolbox (right).

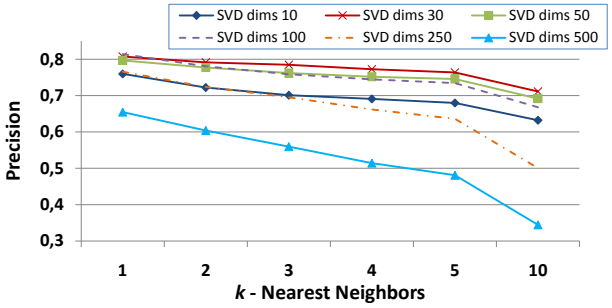

Fig. 2. Dataset A - context, mean precision vs. kNNs vs. SVD dims

The following experiment, Figure 3 aims in providing further insight as to the attained accuracy in computed similarity utilising the features included in the MA-Toolbox using both datasets. Once again, the resulting precision is very low, following the previously mentioned result in Figure 1 (right).

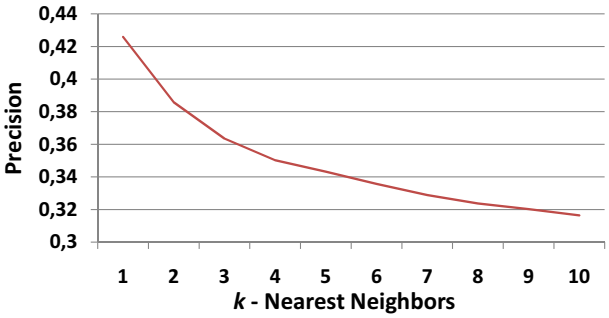

Fig. 3. Dataset A&B - content, mean precision vs. kNNs.

In the next experiment, as shown in Figure 4, we tested the similarity measurement using the contextual information of both dataset A & B. Again, it is

clearly shown that the accuracy of similarity measurement in the tag feature space outperforms similarity in the audio feature space, following the result of Figure 2.

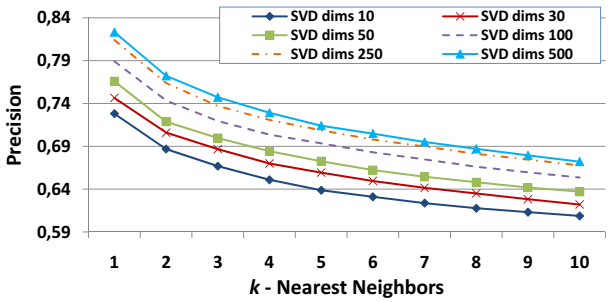

Fig. 4. Dataset A&B - context, mean precision vs. kNNs vs. SVD dimensions using the tracks' metadata.

Finally, we examined the accuracy in similarity measurement using both datasets relying on the contextual in formation of the tracks. The groundtruth in this case is the similarity based on real user listening data from Last.fm. As it can be seen in Figure 5 the contextual information provided by tags offers increased discriminating capability in comparison to the features extracted from the content of the track. Nevertheless, the examined methodology for the calculation of the similarity does not match closely the real user listening data similarity of and thus offering not as high accuracy.

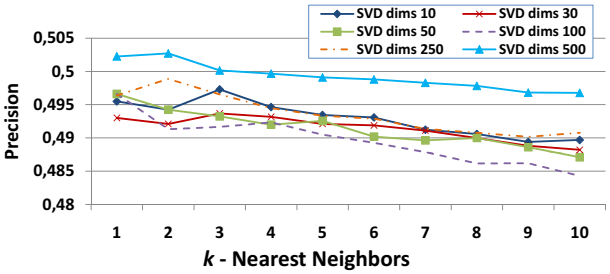

Fig. 5. Dataset A&B - context, mean precision vs. kNNs vs. SVD dimensions using the similarity by Last.fm.

5.3 Discussion

The presented performance evaluation results can be summarised as follows:

– The generic tag-based approach utilised herein outperforms the audio-based method for all k-NN values given the ample amount of tags per track. This

result is in accordance with relevant research stating that the contextual information provided by tags is known to offer increased discriminating capability for the purposes of MIR.
- The similarity measurement methodologies examined herein fail to closely match the real user listening data similarity, providing motivation for techniques that will offer higher accuracy.
- The effect of the SVD dimensionality reduction is of importance to the accuracy of the examined methodology and thus requires tuning.

6 Conclusion

Measuring music similarity is a research area that is of great importance for the purposes of music information retrieval. Different directions exist as to which attributes of a musical datum to retain in order to estimate the similarity between songs. The most common approaches focus on datum metadata, content-based extracted features and "web 2.0" contextual information. Each alternative presents a number of advantages and disadvantages.

In this work, we examine the accuracy of commonly utilised methodologies to musical similarity calculation based on content-based extracted features and "web 2.0" contextual information of the musical data. In addition to common practice groundtruth based on objective metadata we also employ real-life user preference based similarity as provided by Last.fm web service. Experimental results indicate the superiority of the methods based on contextual information and in addition a not close match of these methods to the similarity as perceived by the real-life user preferences.

Future research directions include the examination of more methods that utilise contextual information for musical similarity, experimentation on the number of tags required per musical track in order to establish high accuracy results and the identification of methods that result to a closer match with user perceived similarity.

References

1. Apple: iTunes - Everything you need to be entertained., http://www.apple.com/itunes/
2. Aucouturier, J.J., Pachet, F.: Music similarity measures: What's the use? In: Proc. International Symposium on Music Information Retrieval. pp. 157–1638 (2003)
3. Aucouturier, J.J., Pachet, F.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences 1 (2004)
4. Baumann, S., Hummel, O.: Using cultural metadata for artist recommendations. In: Proc. Web Delivering of Music, International Conference on (2003)
5. Berenzweig, A., Logan, B., Ellis, D.P.W., Whitman, B.P.W.: A large-scale evaluation of acoustic and subjective music-similarity measures. Computer Music Journal 28, 63–76 (2004)
6. Byrd, D.: Organization and searching of musical information, course syllabus (2008), http://www.informatics.indiana.edu/donbyrd/Teach/I545Site-Spring08/SyllabusI545.html

7. Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S.: Using latent semantic analysis to improve information retrieval. In: Proc. Conference on Human Factors in Computing. pp. 281–285 (1988)
8. Karydis, I., Laopodis, V.: Web 2.0 cultural networking. In: Proc. Pan-Hellenic Conference in Informatics (2009)
9. Karydis, I., Nanopoulos, A.: Audio-to-tag mapping: A novel approach for music similarity computation. In: Proc. IEEE International Conference on Multimedia & Expo (2011)
10. Knees, P., Pampalk, E., Widmer, G.: Artist classification with web-based data. In: Proc. International Symposium on Music Information Retrieval. pp. 517–524 (2004)
11. Kontaki, M., Karydis, I., Manolopoulos, Y.: Content-based information retrieval in streaming music. In: Proc. Pan-Hellenic Conference in Informatics. pp. 249–259 (2007)
12. Lamere, P.: Social tagging and music information retrieval. Journal of New Music Research 37, 101–114 (2008)
13. Last.fm: Listen to internet radio and the largest music catalogue online, http://www.last.fm
14. Levy, M., Sandler, M.: Learning latent semantic models for music from social tags. Journal of New Music Research 37(2), 137–150 (2008)
15. Levy, M., Sandler, M.: Music information retrieval using social tags and audio. IEEE Transactions on Multimedia 11, 383–395 (April 2009)
16. Logan, B., Ellis, D.P.W., Berenzweig, A.: Toward evaluation techniques for music similarity. Proc. International Conference on Multimedia & Expo 2003 (2003)
17. McEnnis, D., McKay, C., Fujinaga, I.: jAudio: A feature extraction library. In: Proc. International Conference on Music Information Retrieval (2005)
18. McEnnis, D., McKay, C., Fujinaga, I.: jAudio: Additions and improvements. In: Proc. International Conference on Music Information Retrieval. p. 385 (2006)
19. McFee, B., Barrington, L., Lanckriet, G.: Learning similarity from collaborative filters. In: International Society of Music Information Retrieval Conference. pp. 345–350 (2010)
20. MIREX: Music Information Retrieval Evaluation eXchange
21. Mitrovic, D., Zeppelzauer, M., Breiteneder, C.: Features for content-based audio retrieval. In: Advances in Computers: Improving the Web, vol. 78, pp. 71 – 150. Elsevier (2010)
22. Pampalk, E.: Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns. In: Proc. International Symposium on Music Information Retrieval (2006)
23. Pampalk, E.: Computational Models of Music Similarity and their Application in Music Information Retrieval. Ph.D. thesis, Vienna University of Technology, Vienna, Austria (2006)
24. Pampalk, E., Dixon, S., Widmer, G.: On the evaluation of perceptual similarity measures for music. In: Proc. International Conference on Digital Audio Effects. pp. 7–12 (2003)
25. Pampalk, E.: MA Toolbox, http://www.pampalk.at/ma/
26. Porter, M.F.: The porter stemming algorithm, http://tartarus.org/ martin/PorterStemmer/
27. Slaney, M., Weinberger, K., White, W.: Learning a metric for music similarity. In: Proc. International Conference on Music Information Retrieval. pp. 313–318 (2008)
28. Wang, D., Li, T., Ogihara, M.: 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. pp. 57–62 (2010)