Neurocomputing 107 (2013) 69–76

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Comparing content and context based similarity for musical data

Ioannis Karydis [a,*], Katia Lida Kermanidis [a], Spyros Sioutas [a], Lazaros Iliadis [b]

[a] Department of Informatics, Ionian University, Kerkyra 49100, Greece
[b] Department of Forestry and Management of the Environment and Natural Resources, Democritus University of Thrace, Pandazidou 193, Orestiada 68200, Greece

ARTICLE INFO

Available online 23 October 2012

Keywords:
Music information retrieval
Musical similarity
Extracted musical features
Social network tags
Neural networks

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 data descriptors of limited semantic value, 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, a large scale (20k real music data) similarity measurement is performed using mainstream off-the-shelf methodologies relying on both content and context. In addition, the accuracy of the examined methodologies is tested against not only objective metadata but also real-life user listening data as well. Experimental results illustrate the conditionally substantial gains of the context-based methodologies and not a so close match of these methods with the similarity based on real-user listening data.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

For a classic rock music lover, Led Zeppelin's "Kashmir" and Deep Purple's "Perfect Strangers", may be two similar songs while for a hip-hop admirer the very same songs may be deemed 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 [37,12].

Despite the inherent difficulties in assessing musical similarity, its function is of high value to numerous areas of music information retrieval (MIR) [12]. Based on music-similarity measures [12]: (a) listeners can query using already performed or hummed musical 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 characteristic 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's title, the singer's, and the composer's name or even 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 [12].

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 [8,30]. On the contrary, in automated computer systems the equivalent process based on features extracted from content is much more difficult as the attributes expressed by the extracted features are of very little or lacking any semantic meaning [30]. Moreover, the performance of the widely used and accepted content-based music similarity based on content's global timbre quality, has been reported to reach a limit that is characterised as "glass ceiling" [3,31,12].

On the other hand, contextual knowledge for the purposes of MIR, is derived from numerous sources the most prominent of which is the assignment of information to music through the practice of appointing free-form text (a.k.a. tags) on musical data on the web and the highly popular social media. Based on the previously mentioned ability of the human brain to utilise contextual information for music similarity and the contextually rich semantic nature of the human-generated information that is assigned to the musical works, the important role of tagging in

* Corresponding author.
E-mail addresses: karydis@ionio.gr (I. Karydis), kerman@ionio.gr (K. Lida Kermanidis), sioutas@ionio.gr (S. Sioutas), liliadis@fmenr.duth.gr (L. Iliadis).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.05.033

MIR comes as no surprise. Consequently, measurements of musical similarity based on tags are in cases [25,12,40] 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 have been shown [15] to burden its use in MIR.

To complicate matters further, setting aside the source of information on which the similarity is to be calculated, the requirement of relevance judgments (a.k.a. ground truth) in order to evaluate the calculated similarity is hard to meet due to the previously mentioned subjective character of the musical similarity. In fact it is the same problem that the information retrieval (IR) evaluation field faced since its early beginning due to the "subjectivity in the very concept of relevance" [39]. In the MIR field, the notion of genre offers a debated [27] classification that has been used for evaluation [6,12,35], "assuming that very similar tracks belong to the same genre" [35]. In addition, latest developments concerning the accumulated user assigned labels on musical data over internet social networks have offered an interesting additional alternative to ground truth using analysis of real-user listening patterns [18].

1.1. Motivation and contribution

Bearing in mind the aforementioned importance of music similarity computation, current research has utilised numerous content-based methodologies, the performance of which has been shown to be reaching an upper bound far from the best possible and in cases with little possibility of results' generalisation. Moreover, latest advances in the social media domain have offered the possibility to utilise collectively assigned contextual information on very large musical collections in order to measure similarity and additionally record the listening preferences of users providing, thus, form an alternative to metadata based ground truth.

Accordingly, this paper compares and evaluates content-based versus context-based approaches for measuring music similarity. The contribution of this work is summarised as follows:

- Execution of a large scale, real-data (approx. 20k tracks) similarity measurement.
- Application of broad spectrum off-the-shelf methodologies in order to avoid highly optimised solutions that potentially fit the data under examination.
- Use of both content and contextual information of the data.
- Measure the accuracy of the examined methodologies against not only objective metadata but real-life user listening data as well.

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 concerning both the content- and context-based sources examined herein. Subsequently, Section 4 presents and discusses the experimentation and results obtained, while the paper is concluded in Section 5.

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 [21,31,32,37,3,2,5] due to its importance to numerous areas of MIR.

Content-based similarity has been the cornerstone of automated similarity measurement method in MIR and most research [34,21,31,3,2,5,14] 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 more accurate similarity measurements. The work of Pampalk [31,33] on single Gaussian combined, as submitted to the MIREX 2006 [29] is of high importance as it achieved very high score and in addition, in the 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. [23,24] 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 [4,13] 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 [11] by simply assigning informational tags on musical data. The increasing appeal of the tagging process led to the assignment of large amounts of such information on everything musical. Accordingly, research [15,19,20] expanded in this direction in order to measure the similarity of musical content. Lamere [15] explored the use of tags in MIR as well as issues and possible future research directions for tags while, Levy and Sandler [19] presented a number of information retrieval models for music collections based on social tags. Finally, Wang et al. [40], in contrast to the similarity and classification experiments presented herein, present a performance comparison between content- and context-based features for artistic style clustering that conclude the superiority of context for the purposes of musical style clustering.

The application of artificial neural networks (ANNs) for the purposes of classification in MIR has seen a number of works [26,28,7]. Their extended use in content-based classification is due on the capability of ANNs to "simulate sophisticated logical relationships between features" [26] despite the expensive training required in order to do so. The most common instance of ANNs is the feedforward ANN during which the construction of a network takes place that includes the input, hidden, and output units. Units are interconnected by weights that induce their respective input. The resulting output is then propagated to a transfer function through units to the output unit. The "learning" procedure of ANNs refers to their ability to solve the task given in optimal sense by use of iterative application and weight modification in order to minimise error within a threshold. The backpropagation version of the feedforward ANNs refers to such networks that have no connections that loop but produce a backwards propagation of information based on which the weights of units are updated in order to ameliorate the "learning" procedure.

Genre classification has been used for music classification long before the use of computers as well as in contemporary MIR research [9,38,26]. Nevertheless, its usefulness has been under debate with the key argument of those against being its "limited utility as a goal in itself because of the ambiguities and subjectivity inherent to genre" as stated by McKay et al. [27]. In the same work, a number of proposals concerning the amelioration of the efficiency and effectiveness of the automated genre classifiers are presented, most notably of which, the use of fuzzy logic in order to allow for a musical piece to be member of more than one genre classes and additionally apply weights of importance on each membership.

3. Musical similarity

3.1. Content-based similarity

Content-based approaches assume that musical objects are described by a set of features extracted directly from the content of a musical document [14]. Accordingly, MIR processes depend heavily on the quality of the extracted audio features [24]. In other words, the performance of a classifier or distance metric is strongly defined by the quality of the extracted features. Thus, features with poor expressive capability 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 [24].

In the analysis presented herein, experimentation is done with three known alternatives: (a) content feature extraction based on the jAudio application [23] that produces a set of generic features, (b) the more MIR specific single Gaussian combined method, as implemented in the MA-Toolbox Matlab library [33], that was shown to perform more than adequately in the MIREX contests, and (c) also the MIR oriented MIRtoolbox [17] library.

3.1.1. Generic features

As the generic features can describe a wealth of information for the original data, a large array of such features has been created and maintained for the purposes of this work. 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 [24]. It eliminates the need for re-implementing existing feature extraction algorithms and provides a framework that facilitates the development and deployment of new features [24].

jAudio is able to extract numerous basic features [23]. These features may be one-dimensional (e.g., RMS), or may consist of multi-dimensional vectors (e.g., MFCC coefficients) [24]. Metafeatures are feature templates that automatically produce new features from existing features [24]. These new features function just like normal features-producing output on a per-window basis [24]. Metafeatures can also be chained together. jAudio provides three basic metafeature classes (mean, standard deviation, and derivative).

For the purposes of the experimentation the following features are retained: 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 13 MFCC coefficients, first 10 LPC coefficients, and first five method of moments coefficients.

Accordingly, two mainstream distance measures have been utilised: the Euclidean and the cosine distance.

3.1.2. Targeted features

In order to proceed to the extraction of targeted features, the feature extraction process utilised is based on the single Gaussian combined (G1C) [32]. 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 [31]. 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 [31]. The FPs describe the modulation of the loudness amplitudes per frequency bands, while to some extent can also 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 [31]. 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 [32].

In addition, the feature extraction process and respective similarity calculation offered by the MIRtoolbox [16] have also been utilised. Therein, the distance between audio files is measured based on MFCCs and a selection of distance metrics, having the cosine distance as the default option.

3.2. Context-based similarity

As far as contextual information is concerned, as tags are free-form text assigned by users, pre-processing is a necessity. Tags are initially processed to remove common English language stop words, that is, words that have very low contribution in terms of meaning and then stemmed with the Porter algorithm [36], in order to remove the commoner morphological and in flexional endings from English word aiming at "term normalisation". Accordingly latent semantic analysis (LSA) [10] is employed, in order to alleviate the problem of finding relevant musical data from search tags [15]. 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. Initially, the term-document matrix (TDM) is created detailing the number of times a tag has been assigned at a track. Then, the respective term frequency–inverse document frequency (TFIDF) representation is calculated where TDM counts are modified so that tags with rare assignment are weighted more heavily than common tags using Eq. (1), where $N_{i,j}$ is the number of times the tag i has been assigned to track j, N_j is the number of total tags assigned to track j, D is the number of tracks, and D_i is the number of tracks at which tag i has been assigned

$$TFIDF(i,j) = (N_{i,j}/N_j) * \log(D/D_i) \quad (1)$$

In the final step, singular value decomposition (SVD) is used in order to produce a reduced dimensional representation of the TFIDF matrix that emphasises the strongest relationships and reduces noise using a varying number of dimensions aiming at testing the effect of the SVD approach at the received precision.

3.3. ANN-based classification

As the notion of genre has been previously discussed (Section 2) to offer a useful description of musical classification [27] as well as being interrelated with the notion of similarity, the field of artificial neural networks genre classifiers is of particular interest in this research. The generic features described in Section 3.1.1 have been utilised as observations fed into a feed-forward back-propagation ANN creating thus a pattern recognition network aiming in classifying the given observations into their respective genres. In order to test a widely accessible off-the-shelf such methodology, the neural network toolbox of Matlab [22] has been used.

4. Performance evaluation

In this section we experimentally compare the accuracy of the content and context-based methods using as ground truth both the metadata of the tracks and the similarity provided Last.fm [18] web service based on real-life user listening data. Initially,

the experimental set-up is described, then the results are presented and finally a short summarisation of the key findings is discussed.

4.1. Experimental set-up

For the purposes of performance evaluation of the alternative methods to measure similarity two datasets have been accumulated 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 50 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 5460 discrete tracks for dataset A[1] 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 s clip of the original audio, an audio length commonly considered in related research [40,29].
- *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 has a very large number of tags per track, 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 the following experimentation the focus on is genre information, which is commonly used for evaluating similarity measures [29,12].

As far as the audio content data is concerned, the representation of tracks in the experimentation is based on the following three schemes: (a) Generic content-based 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 13 MFCC coefficients, first 10 LPC coefficients and first five method of moments coefficients, as described in Section 3.1.1. Extraction was achieved using the *jAudio* [24] application for each entire musical datum producing thus a single content feature point of 39 dimensions per track. (b) MA-Toolbox content-based features: single Gaussian combined (G1C) as described in Section 3.1.2. Extraction was achieved through *MA-Toolbox*, a collection of Matlab functions that implement G1C, as described in [32]. (c) MIRtoolbox content-based features: MFCC's based features. Extraction was achieved through *MIRtoolbox*, a collection of Matlab functions, as described in [16].

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 by Porter [36]. 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 the term-document matrix that emphasises the strongest relationships and discards noise.

In the ANN experimentation, a feed-forward backpropagation ANN with one hidden layer has been created containing a varying number of neurons in order to test the effect of the neuron availability. In addition, the experimentation has also included the division of the dataset into training, validation of generality, and testing subsets. In all experiments with the ANN presented herein only dataset A has been used, while evaluation of the performance was only based on the testing subset. The learning function used was the Levenberg–Marquardt backpropagation function, the output layer transfer function was the hyperbolic tangent sigmoid transfer function while the performance function was the mean squared error (MSE) performance function as all implemented in Matlab software. For each of the parameters examined (number of neurons and division of the dataset) the resulting performance of the ANN was averaged over 100 runs due to the randomness in the division of the dataset into training, validation, and testing subsets and in order to receive high quality results.

Initially, the methodologies examined herein have been tested using solely content-based features. Accordingly, Figs. 1–5 report results on similarity measurement accuracy where musical data are represented using features extracted directly from the content of each musical datum. On the other hand, Figs. 6–8 present results where musical data are represented through their respective contextual information. The incorporation of dataset B into some of the experiments for the similarity measurement process aims in using the similarity results of Last.fm as a ground truth. Thus, the intuitive result of using real user listening data as a ground truth similarity is to observe the capability of the examined methodologies to measure similarity in the manner real life users would.

In the evaluation of the similarity between tracks, the precision resulting from the k nearest-neighbours (k-NN) of a query song has been used, i.e., for each query song the fraction of its k-NN that shares the same genre with the query song is measured. In the cases that employ both datasets A and B, queries are selected from dataset A while similar matches are retrieved from both datasets. For the evaluation of the ANN due to the large number of genres (classes) of the data no confusion matrix is presented herein. Evaluation is performed on the ability of the ANN to correctly assign each musical datum of the testing set in the ground truth genres by calculation of the mean precision of the resulting classification for varying number of neurons and different divisions of the dataset into training, validation, and testing subsets.

4.2. Experimental results

In the first experiment, Fig. 1, the accuracy of similarity measurement using solely the content of tracks from subset A has been tested. For this experiment we utilised the generic features extracted using the jAudio application representing thus each track by a 39 dimension vector, while the distance metric under test was the Euclidean and the cosine distance. 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.

[1] Track metadata available at http://goo.gl/hjNmt, last visited on March 3rd, 2012.

In the next experiment, the attained accuracy in computed similarity utilising the features and distance measurement included in the MA-Toolbox has been examined. Fig. 2 presents the resulting precision for varying k number of nearest-neighbours using the G1C features on dataset A. The default setting accuracy provided by the MA-Toolbox is comparable to the accuracy provided by the generic set of features.

The following experiment, Fig. 3 (left) 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 comparable with the precision of the generic features and the MA-Toolbox applied on solely dataset A, following the previously mentioned results in Figs. 1 and 2. Moreover, Fig. 3 (right) shows the results attained for the previous experiment extended to datasets A and B but using as a ground truth the similarity based on real user listening data from Last.fm. As it can be seen, the MA-Toolbox does not manage to closely capture the real user listening data similarity as the precision of the results is lower than the case of having genre as ground truth.

Continuing further, the next experiment examined the attained accuracy in computed similarity utilising the features and distance measurement included in the MIRtoolbox library. Fig. 4 (left) presents the resulting precision for varying k number of nearest-neighbours on dataset A. The default settings accuracy provided by the MIRtoolbox is comparable to the accuracy provided by the generic set of features as well as the MA-Toolbox. Moreover, Fig. 4 (right) shows the results attained for the previous experiment extended to datasets A and B but using as a ground truth the similarity based on real user listening data from Last.fm. As in the case of the MA-Toolbox in Fig. 3 (right), the MIRtoolbox does not also manage to closely capture the real user listening data similarity as the precision of the results is lower than the case of having genre as ground truth.

As final experiment using the content-based features for the representation of the musical data, a feed-forward backpropagation ANN with one hidden layer has been examined. In this case, the ANN is performing a classification task of the input data into genres. Accordingly, the resulting output is comparable with the rest of the experiments that use the genre as a ground truth. Fig. 5 presents the attained accuracy of the ANN, for the aforementioned classification task, for varying number of hidden neurons as well as different divisions of the dataset into training, validation, and testing subsets. In this experiment, the mean MSE value for each run of the validation was approximately 0.02. As it can be seen, the result of the division of the subset, for the values tested, has very little effect in the precision of the classification. On the other hand the number of the hidden neurons does have a significant effect. In addition to the values of number of the hidden neurons presented herein, values up to 5000 neurons have also been under experimentation that showed a further decline in accuracy of classification and thus are not presented. The accuracy attained with the ANN classification is, once again, comparable to the accuracy provided by the generic feature set, the MA-Toolbox and the MIRtoolbox.

Moving on to the similarity measurement using the contextual information, the next experiment presents the accuracy of the similarity measurement using the contextual information of the dataset A tracks. Fig. 6 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.

In the next experiment, as shown in Fig. 7, we tested the similarity measurement using the contextual information of both dataset A and 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 Fig. 6.

Finally, the accuracy in similarity measurement using both datasets relying on the contextual information of the tracks has been examined. The ground truth in this case is the similarity based on real user listening data from Last.fm. As it can be seen in Fig. 8 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 and thus offering not as high accuracy.

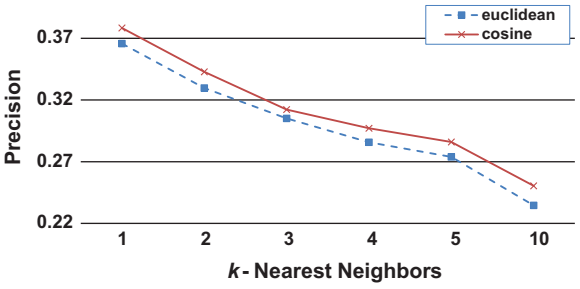


Fig. 1. Generic features, dataset A, mean precision vs. kNNs.

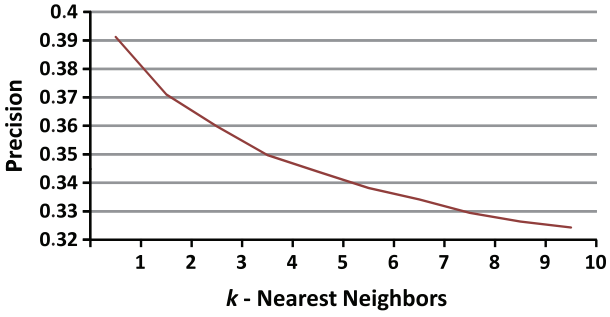


Fig. 2. MA-Toolbox features, dataset A, mean precision vs. kNNs.

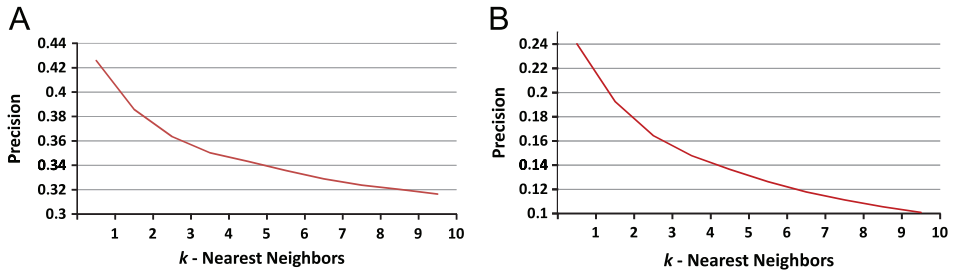


Fig. 3. MA-Toolbox features, datasets A and B, mean precision vs. kNNs: Using genre (left) and the similarity by Last.fm (right) as ground truth.

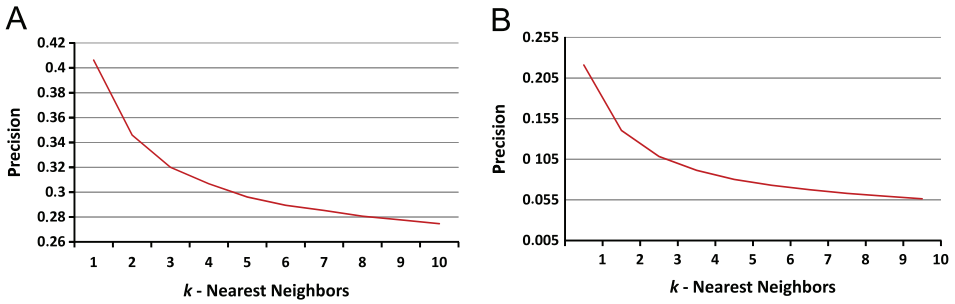


Fig. 4. MIRtoolbox features, mean precision vs. kNNs: dataset A using genre (left) and datasets A and B using the similarity by Last.fm (right) as ground truth.

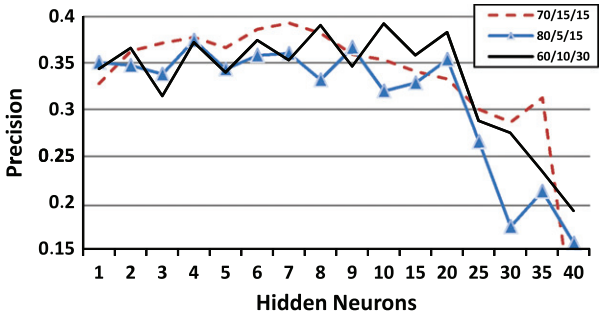


Fig. 5. Generic features, dataset A, mean precision vs. hidden neurons vs. different random division of dataset A into training, validation, and testing subsets (percentages).

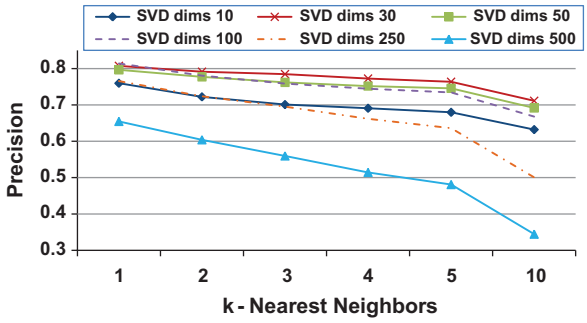


Fig. 6. Contextual data, dataset A, mean precision vs. kNNs vs. SVD dims.

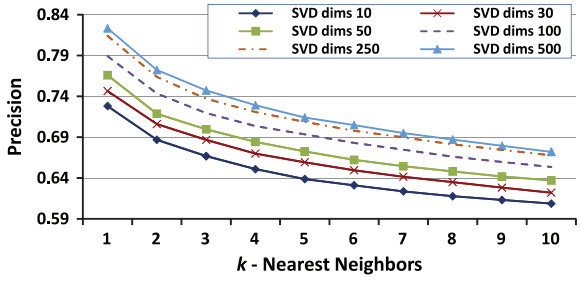


Fig. 7. Contextual data, dataset A and B, mean precision vs. kNNs vs. SVD dimensions using the tracks' metadata.

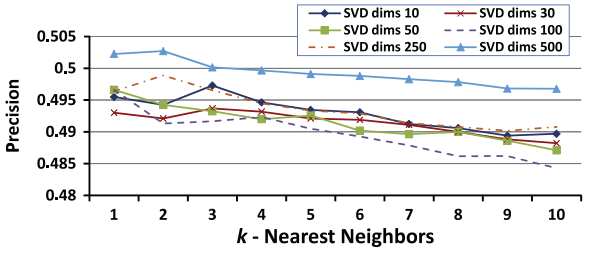


Fig. 8. Contextual data, dataset A and B, mean precision vs. kNNs vs. SVD dimensions using the similarity by Last.fm.

Table 1
Maximum mean precision for each method vs. ground truth type.

Source of information	Similarity measurement approach	Max mean precision	
		Ground truth: genre	Ground truth: Last.fm
Content—audio extracted features	Euclidean and cosine distance of features	0.37	–
	MA-Toolbox	0.42	0.24
	MIRtoolbox	0.40	0.21
	ANN	0.39	–
Contextual—tags	Latent semantic analysis	0.83	0.50

4.3. Discussion

The performance evaluation results can be summarised as follows:

- The generic context-based approach utilised herein outperforms the content-based method for all *k*-NN values given an ample amount of tags per track and use of genre as a ground truth. This result is in accordance with relevant research stating that the contextual information provided by tags is known to offer, under conditions, increased discriminating capability for the purposes of MIR.
- All similarity measurements, including context-based methodologies, examined 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 in the generic context-based approach utilised herein is of importance to the accuracy of the examined methodology and thus requires tuning.
- The generic features have comparable results with alternative methodologies of widely used libraries using content-based information, for their default settings.

Moreover, Table 1 presents an overview of the results obtained herein by comparing the achieved maximum mean precision for each of the methodologies utilised in comparison to the ground truth used in order to evaluate the results. As the application of Euclidean and cosine distance on content audio extracted features is used as a baseline approach serving mainly as a motivation factor for the identification of methodologies that perform better no experimentation on the ground truth based on real user listening data from Last.fm has been performed. For the ANN

method, again no experimentation on the ground truth based on real user listening data from Last.fm has been performed since the ANN is tested on its ability to correctly assign each musical datum of the testing set in the ground truth objective metadata classes.

5. 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 relative to the datum. In addition, the definition of a ground truth judgement as to the similarity of musical documents is a hard problem to solve, towards which, social networks recording user preferences can significantly contribute.

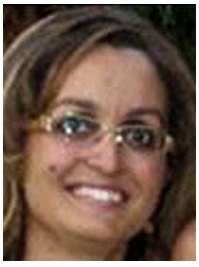
This work examines the accuracy of commonly utilised methodologies to musical similarity calculation based on content-based extracted features and "Web 2.0" contextual information of the musical data. Our work compares results obtained from commonly used distances in MIR research including Euclidean and cosine distance, G1C features and KL, Earth movers distance and broad spectrum off-the-shelf methodologies such as artificial neural networks and latent semantic analysis in order to avoid highly optimised solutions that potentially fit the data under examination. In addition, in contrast to common practice ground truth based on objective metadata in our work we also employ real life user preference-based similarity as provided by Last.fm web service. The comparison of the objective metadata similarity and the real life user preference-based similarity offers an intuitive conditional result concerning the capability of the examined methodologies to measure similarity in the manner real life users would. Experimental results indicate the superiority of the methods based on contextual information and in addition a not so close match of all methods examined 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. 2011. ⟨http://www.apple.com/itunes/⟩.
[2] J.J. Aucouturier, F. Pachet, Music similarity measures: what's the use?, in: Proceedings of the International Symposium on Music Information Retrieval, 2003, pp. 157–1638.
[3] J.J. Aucouturier, F. Pachet, Improving timbre similarity: how high is the sky? J. Negative Results Speech Audio Sci. (2004) 1.
[4] S. Baumann, O. Hummel, Using cultural metadata for artist recommendations, in: Proceedings of the International Conference on Web Delivering of Music, 2003.
[5] A. Berenzweig, B. Logan, D.P.W. Ellis, B.P.W. Whitman, A large-scale evaluation of acoustic and subjective music-similarity measures, Comput. Music J. 28 (2004) 63–76.
[6] C. Boletsis, A. Gratsani, D. Chasanidou, I. Karydis, K. Kermanidis, Comparative analysis of content-based and context-based similarity on musical data, in: Artificial Intelligence Applications and Innovations, 2011.
[7] K. Bozena, Perception-Based Data Processing in Acoustics, Springer-Verlag, GmbH, 2005.
[8] D. Byrd, Organization and searching of musical information, course syllabus, 2008.
[9] R.B. Dannenberg, B. Thom, D. Watson, A machine learning approach to musical style recognition, in: Proceedings of the International Conference on Music Information Retrieval, 1997, pp. 344–347.
[10] S.T. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, Using latent semantic analysis to improve information retrieval, in: Proceedings of the Conference on Human Factors in Computing, 1988, pp. 281–285.
[11] I. Karydis, V. Laopodis, Web 2.0 cultural networking, in: Proceedings of the Pan-Hellenic Conference in Informatics, 2009.
[12] I. Karydis, A. Nanopoulos, Audio-to-tag mapping: a novel approach for music similarity computation, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2011.
[13] P. Knees, E. Pampalk, G. Widmer, Artist classification with web-based data, in: Proceedings of the International Symposium on Music Information Retrieval, 2004, pp. 517–524.
[14] M. Kontaki, I. Karydis, Y. Manolopoulos, Content-based information retrieval in streaming music, in: Proceedings of the Pan-Hellenic Conference in Informatics, 2007, pp. 249–259.
[15] P. Lamere, Social tagging and music information retrieval, J. New Music Res. 37 (2008) 101–114.
[16] O. Lartillot, P. Toiviainen, A matlab toolbox for musical feature extraction from audio, in: Proceedings of the International Conference on Digital Audio Effects, 2007.
[17] O. Lartillot, P. Toiviainen, MIRtoolbox, 2007. ⟨http://www.jyu.fi/music/coe/materials/mirtoolbox⟩.
[18] Last.fm, Listen to internet radio and the largest music catalogue online, 2011. ⟨http://www.last.fm⟩.
[19] M. Levy, M. Sandler, Learning latent semantic models for music from social tags, J. New Music Res. 37 (2008) 137–150.
[20] M. Levy, M. Sandler, Music information retrieval using social tags and audio, IEEE Trans. Multimedia 11 (2009) 383–395.
[21] B. Logan, D.P.W. Ellis, A. Berenzweig, Toward evaluation techniques for music similarity. in: Proceedings of the International Conference on Multimedia and Expo, 2003.
[22] MathWorks, Neural network toolbox for matlab, 2011. ⟨http://www.mathworks.com/products/neural-network/⟩.
[23] D. McEnnis, C. McKay, I. Fujinaga, jAudio: a feature extraction library, in: Proceedings of the International Conference on Music Information Retrieval, 2005.
[24] D. McEnnis, C. McKay, I. Fujinaga, jAudio: additions and improvements, in: Proceedings of the International Conference on Music Information Retrieval, 2006, p. 385.
[25] B. McFee, L. Barrington, G. Lanckriet, Learning similarity from collaborative filters, in: International Society of Music Information Retrieval Conference, 2010, pp. 345–350.
[26] C. McKay, I. Fujinaga, Automatic genre classification using large high-level musical feature sets, in: Proceedings of the International Conference on Music Information Retrieval, 2004, pp. 525–530.
[27] C. McKay, I. Fujinaga, Musical genre classification: is it worth pursuing and how can it be improved?, in: Proceedings of the International Conference on Music Information Retrieval, 2006, pp. 101–106.
[28] A. Meng, P. Ahrendt, J. Larsen, Improving music genre classification by short-time feature integration, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. 497–500.
[29] MIREX, Music information retrieval evaluation exchange, 2011. ⟨http://www.music-ir.org/mirex/wiki/MIREX_HOME⟩.
[30] D. Mitrovic, M. Zeppelzauer, C. Breiteneder, Features for content-based audio retrieval, in: Advances in Computers: Improving the Web, vol. 78, Elsevier, 2010, pp. 71–150.
[31] E. Pampalk, Audio-based music similarity and retrieval: combining a spectral similarity model with information extracted from fluctuation patterns, in: Proceedings of the International Symposium on Music Information Retrieval, 2006.
[32] E. Pampalk, Computational Models of Music Similarity and their Application in Music Information Retrieval, Ph.D. Thesis, Vienna University of Technology, 2006, Vienna, Austria.
[33] E. Pampalk, MA Toolbox. 2007 ⟨http://www.pampalk.at/ma/⟩.
[34] E. Pampalk, S. Dixon, G. Widmer, On the evaluation of perceptual similarity measures for music, in: Proceedings of the International Conference on Digital Audio Effects, 2003, pp. 7–12.
[35] E. Pampalk, A. Flexer, G. Widmer, Improvements of audio-based music similarity and genre classification, in: Proceedings of the International Conference on Music Information Retrieval, 2005, pp. 628–633.
[36] M.F. Porter, An algorithm for suffix stripping, Program 14 (1980) 130–137.
[37] M. Slaney, K. Weinberger, W. White, Learning a metric for music similarity, in: Proceedings of the International Conference on Music Information Retrieval, 2008, pp. 313–318.
[38] G. Tzanetakis, P.R. Cook, Musical genre classification of audio signals, IEEE Trans. Speech Audio Process. 10 (2002) 293–302.
[39] J. Urbano, Information retrieval meta-evaluation: challenges and opportunities in the music domain, in: Proceedings of the International Society for Music Information Retrieval, 2011, pp. 609–614.
[40] D. Wang, T. Li, M. Ogihara, Are tags better than audio? The effect of joint use of tags and audio content features for artistic style clustering, in: Proceedings of the International Society for Music Information Retrieval, 2010, pp. 57–62.

Ioannis Karydis was born in Greece, in 1979. He received a BEng (2000) in Engineering Science and Technology from Brunel University, UK, an MSc (2001) in Advanced Methods in Computer Science from Queen Mary University, UK, and a PhD (2006) in Mining and Retrieval Methods for Acoustic and Symbolic Music Data from the Aristotle University of Thessaloniki, Greece. He has contributed to 28 academic publications and currently is a Contract Lecturer at the Ionian University, Greece. His research interests include music databases, music information retrieval (indexing and searching), musical thumbnailing, cultural information systems, privacy issues in databases, and query processing.



Katia Lida Kermanidis graduated from the Electrical and Computer Engineering Department, University of Patras, in 1999. She received her PhD Diploma in Syntactic Dependencies' Learning from the same department in 2005. Her research interests include artificial intelligence, natural language processing, grammar development, information retrieval, linguistic data mining, ontology extraction. She has participated in European and Greek research projects in the areas of Computational Lexicography, Language and Speech processing, Corpus Processing, and Ontology Extraction. She is a member of the teaching staff of the Department of Informatics, Ionian University, since 2005, and has authored 10 journal articles and over 25 conference papers.

Spyros Sioutas was born in Greece, in 1975. He graduated from the Department of Computer Engineering and Informatics, School of Engineering, University of Patras, in December 1997. He received his PhD degree from the Department of Computer Engineering and Informatics, in 2002. He is now an Assistant Professor in Informatics Department of Ionian University. His research interests include data structures and databases, P2P data management, data warehouses and data mining, computational geometry, GIS and advanced information systems. He has published over 60 papers in various scientific journals and refereed conferences.

Lazaros S Iliadis received a BSc in Mathematics and a PhD in Expert Systems from Aristotle University of Thessaloniki, Greece and an MSc, Master of Science, Computer Science from University of Wales, UK. He is currently an Associate Professor at the department of Forestry and Management of the Environment and Natural Resources, Democritus University of Thrace, Greece. His main research interests include artificial neural networks, fuzzy logic and systems modelling, multiagent systems, expert systems, machine learning, autonomous agents, and decision support systems. He has authored one book and more than 90 publications in international scientific journals and proceeding conferences. He is also Guest Editor in six academic journals.