ANNOTATING MUSIC COLLECTIONS: HOW CONTENT-BASED SIMILARITY HELPS TO PROPAGATE LABELS

Mohamed Sordo, Cyril Laurier, Òscar Celma
Music Technology Group
Universitat Pompeu Fabra
{msordo, claurier, ocelma}@iua.upf.edu

ABSTRACT

In this paper we present a way to annotate music collections by exploiting audio similarity. Similarity is used to propose labels (tags) to yet unlabeled songs, based on the content–based distance between them. The main goal of our work is to ease the process of annotating huge music collections, by using content-based similarity distances as a way to propagate labels among songs.

We present two different experiments. The first one propagates labels that are related with the style of the piece, whereas the second experiment deals with mood labels. On the one hand, our approach shows that using a music collection annotated at 40% with styles, the collection can be automatically annotated up to 78% (that is, 40% already annotated and the rest, 38%, only using propagation), with a recall greater than 0.4. On the other hand, for a smaller music collection annotated at 30% with moods, the collection can be automatically annotated up to 65% (e.g. 30% plus 35% using propagation).

1 INTRODUCTION

Manual annotations of multimedia data is an arduous task, and very time consuming. Automatic annotation methods, normally fine-tuned to reduced domains such as musical instruments or limited to sound effects taxonomies, are not mature enough to label with great detail any possible sound. Yet, in the music domain the annotation becomes more complex due to the time domain frame.

The purpose of making music easily accessible implies a condition of describing music in such a way that machine learning can understand it [1]. Specifically, these two steps must be followed: to build music descriptions which can be easily maintained, and to exploit these descriptions to build efficient music access systems that help users find music in large collections. There are a lot of ways to describe music content, but we can basically classify the descriptors in three groups: editorial meta-data, cultural meta-data, and acoustic meta-data [1].

As a paradigmatic example, the Music Genome Project is a big effort to "capture the essence of music at the fundamental level" by using over 400 attributes to describe songs. To achieve this, more than 40 musicologists have been annotating thousands of files since 2000. Based on this knowledge, a well–known system named Pandora [1] creates playlists by exploiting these human–based annotations. It is clear that helping these musicologists can reduce both time and cost of the annotation task.

Thus, the main goal of our work is to ease the process of annotating music collections, by using content-based similarity distance as a way to propagate labels among songs.

2 RELATED WORK

Nowadays, content-based retrieval systems can not classify, identify and retrieve as well as humans can. This is a common problem in the multimedia field, like in image or video annotation. But in the latter fields many attempts have been made [2][3].

Semantic audio annotation, however, has not been as studied as image or video annotation, except the work by Whitman [4] or Barrington et al. [5][6]. Barrington et al. have made significant advances in semantic annotation of songs for music information retrieval (MIR) using MFCC's to describe music content and HMM's trained on timbre and rhythm for computing similarity between songs. Their idea was basically based on other work that represented image semantic annotation as a supervised multi-classification problem [7] .

In the MIR field, only a few works are dealing with the problem of detecting mood using audio content. Although some results are promising (e.g [8], [9]), there is no standard or clearly defined proposals about the categories and the features to use. In the following experiments, we will check if tags about styles and moods can be propagated using content–based (CB) similarity.

3 EVALUATION

The goal of this paper is to prove empirically how content-based similarity can help to propose labels to yet unlabeled songs, and thus reducing the hard effort of manually annotating songs. For our purpose, the content–based similarity can be seen as a black box. That is to say, given a

[1] http://www.pandora.com

seed song, the module returns a list of the *ith* most similar songs. This study employs a CB module that considers not only timbrical features (e.g. MFCC), but some musical descriptors related to rhythm, tonality, etc. [11].

We present two different experiments. The first one propagates labels that are related with the style of the piece, whereas the second experiment deals with mood labels. The problem with the Magnatune collection is that there is only one human that annotated the tracks, when normally a ground truth of this nature should be pair–reviewed. Yet, we validated a large amount of the annotated songs by listening to them.

3.1 Propagation of music style labels

The ground truth for the style experiment consists of 29 different labels (like *Rock*, *Instrumental*, *Classical*, *Relaxing*, etc.), and 5481 annotated songs.

The evaluation process was the following, for a partially annotated collection (10%,..., 50%), we use the CB module to get the *ith*–similar (i=10, 20 and 30) songs —and their tags— to a given one, to propose tags based on the tags from these similar songs. However, we did not propose those tags that appeared with a frequency less than 20%.

3.1.1 Evaluation metrics

The metrics used to evaluate the styles experiments were initially Precision/Recall and F_2-Measure (giving more weight to Recall). In our case, Recall seems to be more informative since our purpose is to know how well the tags can be propagated. However, neither P nor R take into account the frequencies (i.e. ranking) of the tags obtained from the similar songs. Thus, we used the Spearman's rank correlation coefficient, or Spearman ρ, which is defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

Where d_i represents the distance between each rank of pair of values —in our case labels in the ground truth and labels in the proposed tags— To compute the distances we assume that the frequency of manually annotated labels is equal to 1.

3.1.2 Results

For the style experiment, we ran different configurations and we computed the average metrics. A special case is when using the 100% annotated songs (see the results in Table 1). This experiment is used to test whether the CB similarity is good for propagating labels. There are four different configurations when retrieving the most similar songs to a given one: do not apply any constraint, or filter by artist/album. The constraints, then, are: filtering the similarity results by same Artist, same Album, or by same Artist and Album. The latter case makes only sense when the songs appears in compilations, various artists albums, etc. When filtering by artist or by album we make sure that the most similar songs to a given one are not from the same artist or the same album. That of course decreases the Precision/Recall measure. We can see from the results, that to achieve more precision and recall when applying a constraint, we need to increase the number of similar songs, which makes sense because we are not taking into account similar songs that are closer to a given one.

Sims.	Constraint	P	R	F_2	ρ
10	None	0.56	0.84	0.72	0.51
	Artist	0.41	0.58	0.51	0.23
	Album	0.50	0.71	0.62	0.34
	Artist & Album	0.43	0.59	0.53	0.19
20	None	0.56	0.82	0.71	0.49
	Artist	0.48	0.61	0.56	0.26
	Album	0.53	0.72	0.64	0.35
	Artist & Album	0.48	0.61	0.56	0.24
30	None	0.60	0.77	0.70	0.45
	Artist	0.50	0.58	0.55	0.28
	Album	0.56	0.67	0.63	0.37
	Artist & Album	0.50	0.59	0.55	0.27

Table 1. Experiments with the 100% annotated collection. The Precision/Recall measure, the F_2-measure and the Spearman ρ measure are proportional to the number of similar songs. When constraints are present, these measures decrease.

Now, table 2 shows the results of propagating a partially annotated collection. The Spearman ρ coefficient, as well as Precision/Recall and F_2-measure, grows when increasing the percentage of songs annotated in the collection. Interestingly enough, the values decrease when increasing the number of neighbours (from 10 to 30) for a given song.

Annotation	Sims.	P	R	F_2	ρ
20%	10	0.32	0.29	0.30	0.24
	20	0.22	0.17	0.19	0.16
	30	0.08	0.05	0.06	0.06
40%	10	0.57	0.59	0.58	0.43
	20	0.56	0.52	0.53	0.41
	30	0.49	0.39	0.42	0.34
50%	10	0.61	0.67	0.64	0.47
	20	0.61	0.61	0.61	0.45
	30	0.57	0.51	0.53	0.41

Table 2. Experiments with the 20%, 40% and 50% annotated collection. The Precision, Recall and F_2-measure and the Spearman ρ values grow with a higher percentage of annotated songs, and a smaller number of similar songs.

Finally, we propose another experiment that is to automatically annotate songs in a music collection by means of the propagation process. The results are presented in Table 3. It is clear that the percentage of songs automatically annotated by CB similarity increases when the number of already annotated songs grows. But, we can see

an interesting exception here, that is the 40% annotated collection performs better (up to 38.68% new propagated labels, with a low Recall 0.4) than the 50% one. This could be due to the random process of splitting the ground truth and the test set from the collection. Furthermore, we can see how the percentage of songs automatically annotated is inversely proportional to the number of similar songs used by the CB similarity module (in contrast with the results from the 100% annotated collection, see Table 1, when applying any constraint).

Annot.	Sims.	Propagation with Recall > 0.8	> 0.6	> 0.4
20%	10	17.515%	21.365%	24.977%
	20	8.666%	12.352%	15.453%
	30	2.554%	3.758%	5.145%
40%	10	28.01%	33.46%	38.68%
	20	22.50%	28.92%	34.32%
	30	15.22%	20.82%	26.22%
50%	10	26.77%	31.62%	35.92%
	20	22.66%	28.74%	33.37%
	30	17.48%	23.15%	28.44%

Table 3. Extending annotations of a music collection by means of CB similarity. We observe that the propagation grows with a smaller number of similars and a higher percentage of annotated songs, except for the case of 40% and 50%.

3.2 Propagation of mood labels

For the moods experiment, the first issue is the choice of the taxonomy. As advised by Juslin et al. in [10], in order to make our experiment and to build a ground truth that achieve the best agreement between people, we should consider few categories. We used a reduced version of the Magnatune online library. This collection offers a set of playlists based on mood [2]. We clustered the 150 mood playlists to fit in our few categories paradigm. The adjectives proposed by Juslin: happiness, sadness, anger and fear in [10] have been applied by Feng et al. in [9] and proved to give satisfying results. As the collection is mostly focused on popular and classical music, the "fear" adjective has been extended to a larger category called "mysterious". Using Wordnet [3] we have joined the possible playlists together in the following four categories : happy, sad, angry and mysterious. Then, a listener was asked to validate each song label. We obtained a ground truth database of 191 songs with the distribution in mood shown in Table 4. For each song, there is only one mood label. It is not an equal distribution but there is enough data in each category to experiment with the CB similarity.

[2] http://www.magnatune.com/moods/

[3] http://wordnet.princeton.edu/

Mood	Happy	Sad	Angry	Mysterious
Songs	67	61	34	29

Table 4. Mood distribution of the ground truth

GT/Predicted	Angry	Happy	Mysterious	Sad
Angry	27	7	1	1
Happy	4	55	1	2
Mysterious	8	6	7	5
Sad	4	16	2	35

Table 5. Confusion matrix for the mood experiment with a 100% annotated collection.

3.2.1 Evaluation metrics

To evaluate the mood results, we used two measures. First we wanted to check if the system was able to guess the correct mood label (there is only one possible label per song). We evaluated the Precision just considering the first result using Precision at 1, also called P@1.

$$P@1 = \begin{cases} 1, & best\,proposed\,label = real\,label \\ 0, & otherwise \end{cases} \quad (2)$$

We averaged this value over all the examples. This metric helps us to understand if the system can predict the correct mood label. However it does not take into account the relative frequencies. Then another measure would be needed to evaluate this aspect. We weighted the frequencies of the proposed label and normalized to compute a weighted Precision at 1, that we will call wP@1. It is equal to the frequency value of the correct label over the sum of all the proposed label frequencies:

$$wP@1 = \frac{freq.\,correct\,label}{\sum freq.\,proposed\,labels} \quad (3)$$

3.2.2 Results

To have an overview of the system performance for each mood, we built a confusion matrix in Table 5. It has been computed using 100% of the collection annotated. Each row gives the predicted mood distribution (considering only the best label) for each mood in the ground truth. Looking at the confusion matrix we observe that a CB similarity approach can propagate relatively well the "happy", "angry", and "sad" labels. However the "mysterious" label does not give good results. We can explain this by the fact that it might be the most ambiguous concept of these categories. Table 6 presents the average P@1 and wP@1 values per mood.

	Angry	Happy	Mysterious	Sad	All
P@1	0.72	0.89	0.27	0.61	0.62
wP@1	0.65	0.62	0.22	0.59	0.52

Table 6. P@1 and wP@1 values averaged for each mood

It confirms what we have in the confusion matrix, the "happy" category gives the best result. However looking at the values of wP@1, we note that if "happy" is the most guessed mood, the system gives more reliability to its results about the label "angry".

In our last experiment we wanted to evaluate how well the mood labels can be propagated if we annotate just partially the collection. We computed the P@1 for 70%, 50% and 30% of the database and obtain the results written in Table 7. It shows that for 30% of the collection annotated, the system can propagate correctly the tags up to 65% of the collection.

Initial annotation	70%	50%	30%
P@1	0.60	0.44	0.5
Correctly annotated after prop.	88%	72%	65%

Table 7. Evaluation of the mood label propagation with the initially percentage of annotated songs.

As the CB approach may not consider important aspects that can infer the mood, all these performances should be improved by using dedicated descriptors and approach or meta-data, like information about the title, the style or the lyrics.

4 CONCLUSIONS AND FUTURE WORK

Our objective was to test how the content–based similarity can propagate labels. For styles, we have shown that with a 40% annotated collection, we can reach a 78% (40%+38%) annotated collection with a recall greater than 0.4, only using content–based similarity. In the case of moods, with a 30% annotated collection we can automatically propagate up to 65% (30% +35%). These results are quite encouraging as content–based similarity can propagate styles and moods in a surprisingly effective manner. Of course there are some limitations as the example of the "mysterious" label, the concept has to be clearly encoded in the music for the content–based propagation to work. For the moods we will try to experiment with a larger database, different taxonomies and more concepts. With our current mood results it may not be possible to generalize but it shows the potential of the technique. In general, to enhance the performance of such an automatic annotation system we would use a hybrid approach combining content–based, user feedback and social networks informations. But as shown by the satisfying results, our propagation system based on content–based similarity would already ease a lot the annotation process of huge music collections.

5 ACKNOWLEDGEMENTS

This research has been partially supported by the e-Content plus project VARIAZIONI [4] . We are also very grateful for the help of Jens Grivolla and Joan Serrà, and their advice about evaluation metrics.

[4] http://www.variazioniproject.com

6 REFERENCES

[1] Pachet, F. "Knowledge Management and Musical Metadata", *Encyclopedia of Knowledge Management*.

[2] Jeon, J. and Lavrenko, V. and Manmatha, R. "Automatic image annotation and retrieval using cross-media relevance models", *26th ACM SIGIR conference on Research and development in information retrieval pages 119-126*, Toronto, Canada, 2003.

[3] Wenyin, L. and Dumais, S. and Sun, Y. and Zhang, H. and Czerwinski, M. and Field, B. "Semi-automatic image annotation", *INTERACT2001, 8th IFIP TC, volume 13, pages 9–13*, 2001.

[4] Whitman, B.A. "Learning the meaning of music", *PhD Thesis*, Massachusetts Institute of Technology, 2005.

[5] Barrington, L. and Chan, A. and Turnbull, D. and Lanckriet, G. "Audio Information Retrieval Using Semantic Simlarity", *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Hawaii, 2007.

[6] Turnbull, D. and Barrington, L. and Torres, D. and Lanckriet, G. "Exploring the Semantic Annotation and Retrieval of Sound", *CAL Technical Report CAL-2007-01*, San Diego, 2007.

[7] Carneiro, G. and Vasconcelos, N. "Formulating Semantic Image Annotation as a Supervised Learning Problem", *Computer Vision and Pattern Recognition, volume 2*, IEEE Computer Society Conference, San Diego, 2005.

[8] Lu,L Liu,D Zhang H.J. "Automatic mood detection and tracking of music audio signals", *IEEE transactions on audio, speech and language processing, volume 14, pages 5-18*, 2006

[9] Feng, Y. and Zhuang, Y. and Pan, Y. "Music Information Retrieval by Detecting Mood via Computational Media Aesthetics", *IEEE/WIC International Conference on Web Intelligence*, Washington DC, 2003.

[10] Juslin, P.N. and Sloboda, J.A. *Music and Emotion: Theory and Research*. Oxford University Press, 2001.

[11] Cano, P. et al. "An Industrial-Strength Content-based Music Recommendation System. *28th ACM SIGIR Conference*, Salvador, Brazil, 2005.