From Improved Auto-taggers to Improved Music Similarity Measures

Klaus Seyerlehner[1], Markus Schedl[1], Reinhard Sonnleitner[1],
David Hauger[1], and Bogdan Ionescu[2]

[1] Johannes Kepler University
Department of Computational Perception
Linz, Austria
[2] University Politehnica of Bucharest
Image Processing and Analysis Laboratory
Bucharest, Romania

Abstract. This paper focuses on the relation between automatic tag prediction and music similarity. Intuitively music similarity measures based on auto-tags should profit from the improvement of the quality of the underlying audio tag predictors. We present classification experiments that verify this claim. Our results suggest a straight forward way to further improve content-based music similarity measures by improving the underlying auto-taggers.

Keywords: music information retrieval, music similarity, auto-tagging, music recommendation, tag prediction

1 Introduction

Audio *tags* are semantic textual annotations (e.g., *"beat"*, *"fast"* or *"rock"*) that are used to describe songs. Typically, tags are collected by large online music platforms such as *Last.fm*[3] that allow users to annotate the songs they are listening to. However, there also exist several other methods to collect tag information [13]. Audio tags can also be obtained through surveys, music annotation games or web-mining. Another variant which is in the focus of this paper is to obtain tag information via *auto-tagging*. An auto-tagger is typically a purely content-based method (i.e. only based on a set of audio features extracted out of the audio signal) for predicting tags which might be associated with a song. Consequently, one can interpret an auto-tagger as a method that transforms *an audio feature space* into *a semantic space*, where music is described by words. This process is often referred to as *automatic tag prediction* or *automatic tag classification*.

While automatic tag classification can be viewed just as an interesting performance task that extends traditional genre classification to multi-label classification, there also exist several application scenarios where auto-tagging can

[3] www.last.fm

be extremely beneficial. For example, auto-tags can be used to visualize and explore music collections in a semantic space without relying on community data, which is typically incomplete or unavailable for some songs in a personal music collection. Another application field of auto-tags is to compute song similarities from automatically estimated tags. This approach to music similarity is of special interest in this paper, as we want to study how the quality of the estimated auto-tags influences the quality of a music similarity measure that is build on top of them. Intuitively, a tag based music similarity measure should profit from improving the quality of the underlying tag predictors. However, no empirical evidence of this assumption exists to the best of our knowledge. Thus, the main contribution of this paper is to fill this gap and provide experimental evidence of this relation. From a technical point of view, this relation is especially interesting, as it transforms the ill-defined task of improving a music similarity measure into the well-defined machine learning task of predicting audio tags and defines a straightforward way to improve content-based music similarity measures.

The outline of the remainder of this paper is as follows: In the subsequent section we give a brief introduction to automatic tag classification, as well as auto-tag based music similarity and discuss related work in these research areas. Then, in Section 3 we present the outline of the conducted experiment, which is structured into two successive experiments. We report on these sub-experiments in Sections 4 and Section 5 respectively. Finally, in Section 6 we provide a brief summary and discuss the obtained results.

2 Related Work

While automatic tag prediction recently gained a lot of research attention and can be considered an emerging research area in Music Information Retrieval (MIR), the idea of predicting tags is relatively old. To the best of our knowledge Whitman et al. [17], Slaney [12] and Berenzweig [1] were the first to introduce concepts related to auto-tags. While Slaney was working on animal sounds only, he already introduced the concept of an acoustic and a semantic space. In contrast, Whitman was already working on music and interpreted automatic tag prediction as a multi-label classification problem, while Berenzweig called the semantic tag space *"anchor space"* and was the first to compute similarities among songs based on tag information. However, it seems that in these early days the lack in computational resources and the unavailability of adequate tag sources limited further development in this research direction. Then driven by the general growing interest in tags in the MIR community around the year 2006 this idea was picked up again by West [16], Eck et al. [4], Mandel et al. [7] and Turnbull et al. [14]. The latter introduced in [14] a first formal definition of the tag prediction task:

The task of predicting tags can be interpreted as a special case of multi-label classification and can be defined as follows: Given a set of tags $T = \{t_1, ..., t_A\}$ and a set of songs $S = \{s_1, ..., s_R\}$ predict for each song $s_j \in S$ the tag annotation vector $y = (y_1, ..., y_A)$, where $y_i > 0$ if tag t_i has been associated with the audio

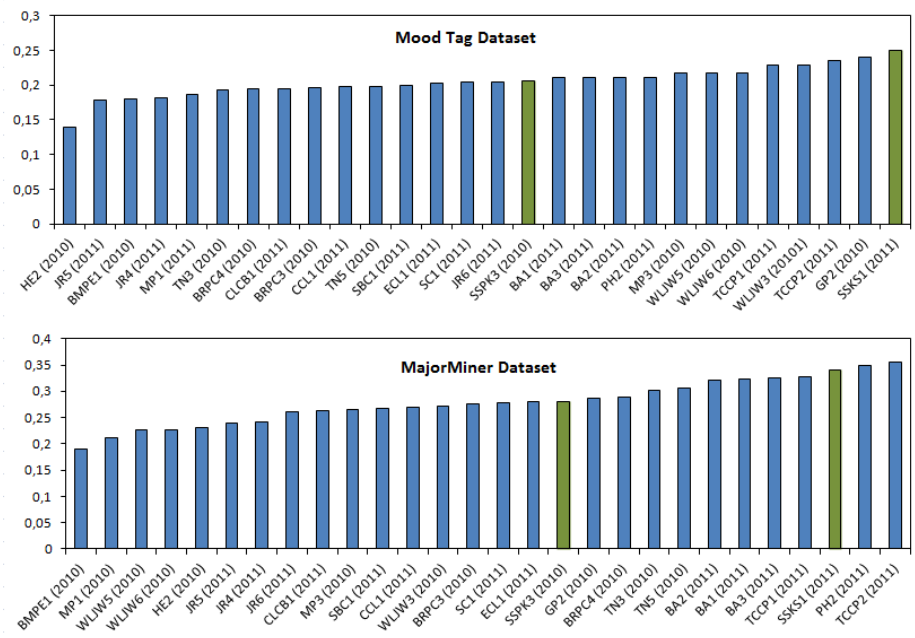

Fig. 1. *Comparison of tag classification algorithms (**average per tag f-Score**) of the MIREX 2010 and MIREX 2011 campaigns.*

track by a number of users, and $y_i = 0$ otherwise. Thus, the y_i's describe the strength of the semantic association between a tag t_i and a song s_j and are called *tag affinities*, *semantic weights* or *tag profiles*. If the semantic weights are mapped to $\{0, 1\}$, then they can be interpreted as class labels, which can be used for training and evaluating tag classifiers.

Recent research work in automatic tag prediction has mainly put the focus on the classification part. For instance in [2] and in [8] the authors propose to extend straightforward binary classification strategies by introducing a second layer of tag classifiers. The inputs of the second classification layer are the predictions of the binary tag classifiers from the first layer. This advanced approach allows to make use of inter-tag correlations in the second classification stage and is one way of improving the classification part of an auto-tagger.

Another recent trend in the context of automatic tag classification is to use auto-tags for music similarity estimation [3, 15]. The common main idea behind *auto-tag based music similarity systems* is to first estimate a song's tag profile and then compare the estimated tag profiles of two songs. Interestingly, during the last two runs of the MIREX Audio Music Similarity and Retrieval task (2010 and 2011), hybrid (content- and tag-based) music similarity measure achieved the first rank [10]. This clearly indicates that auto-tag based similarity will become an important part of music similarity estimation.

3 Experiment: Outline

Intuitively, qualitative improvements in automatic tag classification should indirectly also lead to qualitative improvements of auto-tag based music similarity systems. To experimentally verify this assumption, we have conducted an experiment which is subdivided into two phases:

- In the first phase (*Train Auto-Taggers*) of the experiment, we train two sets of tag classifiers on the same tag classification datasets: A high quality set of tag classifiers (*high*) and a set of low quality tag classifiers (*low*). For both tag classification datasets we ensure that the *high* quality tag classifiers perform significantly better than the *low* quality tag classifiers.
- Then in the second phase (*Estimate Tag Profile Similarities*) of the experiment, we built two tag-based music similarity measure precisely in the same way. The only difference is that, one is based on the *high* quality tag-predictors and the other one is based on the *low* quality tag predictors. Both similarity measures are then evaluated via nearest-neighbor genre classification on six well-known datasets to identify qualitative differences between the two variants. In case the assumption about the relation between tag classification and tag-based music similarity is correct, it is expected that the music similarity measures based on the high quality tag predictors performs significantly better than the similarity measure based on the low quality tag predictors.

The following two sections present details and results of the execution of the two sub-experiments.

4 Phase 1: Train Auto-Taggers

To generate two sets of tag predictors of different quality we make use of the tag classification systems we have submitted to the MIREX tag classification tasks. The low quality tag predictors are generated by our 2010 submission [10], while the high quality tag predictors are generated by the improved submission in 2011 [9]. Both submissions are based on so-called *block-level* audio features [11]. In contrast to standard audio features these features allow to better capture local temporal information and together form a highly descriptive audio feature set. The descriptive power of this feature set has already been demonstrated during several evaluation campaigns (MIREX'2010 [10], MIREX'2011 [9] and MediaEval'2011 [5]).

The main differences between the 2010 submission and the 2011 submission are that two additional block-level features, the *Local Single Gaussian Model* (LSG) and the *George Tzanetakis Model* (GT), were added in 2011 and that the classification method was changed. In the 2010 submission the dimensionality of the high dimensional audio feature space was first reduced via a PCA (the extracted block-level forms a 9448 dimensions vector space), as it was not tractable to use the uncompressed feature set in combination with a *support*

vector machine classifier. For our 2011 submission we decided to directly use the uncompressed feature set and replace the *support vector machine* classifier by a *random forest* classifier, since random forest classifiers can handle very high dimensional feature spaces and can make use of multi-core CPUs. For a more detailed description of the two algorithms we refer to [10] and [9].

Figure 1 visualizes the average per tag f-Score of all submissions in 2010 and 2011. This allows to compare our system to other tag classification approaches, but what is even more interesting in the context of this paper is that there is obviously a significant improvement of our submission in 2011 over our submission in 2010. Thus, this comparison suggests to use the 2010 submission to generate the low quality tags and the 2011 system to generate the high quality tags. It is, however, worth mentioning that the quality of our 2010 system, although it is used to generate the low quality tag predictors, is still quite competitive and not just a baseline system. In the following we introduce two tag classification datasets. These datasets are then used to first experimentally verify the qualitative difference of the two approaches and then to learn a pair of low and high quality tag predictors form each dataset.

4.1 Datasets

Magnatagatune The first dataset in our evaluation is the Magnatagatune [6] dataset. This huge dataset contains 21642 songs annotated with 188 tags. The tags were collected by a music and sound annotation game, the TagATune[4] game. The dataset also includes 30 seconds audio excerpts of all songs that have been annotated by the players of the game. All the tags in the dataset have been verified (i.e. a tag is associated with an audio clip only if it is generated independently by more than 2 players, and only tags that are associated with more than 50 songs are included). From the tag distribution (Figure 2) one can see that in terms of binary decisions (tag present / not present), the classification tasks are extremely skewed. So 110 out of the 188 tags apply to less than 1% of all songs and the 87 most frequently used tags account for 89.86% of all annotations.

RadioTagged The second dataset in our evaluation is called RadioTagged, because the audio files in this dataset were recorded from internet radio streams. Audio fingerprinting was used to identify the recorded tracks and retrieve artist, album, song name and cover art. In a second step *Last.fm* was queried with artist and track name to obtain tags on track level. We only kept those songs for which we were able to retrieve tags. After this process we ended up with a total of 10557 full length songs and 1072 tags. From Figure 3 one can see an effect similar to the one of the Magnatagatune dataset: most of the tags do only apply to a fraction of all songs. This is especially interesting as the two datasets originate from different annotation processes (annotation game and social tagging), but according to their summarization plots the general structure of both datasets is

[4] http://www.tagatune.org

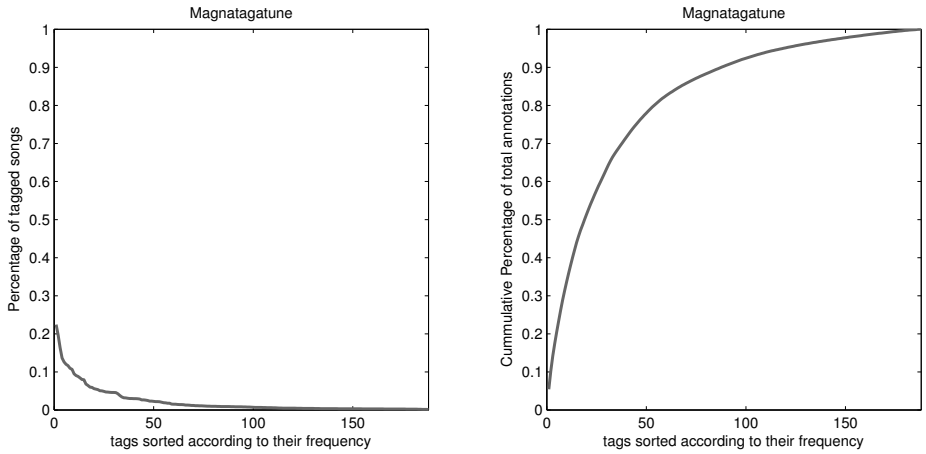

Fig. 2. *Percentage of annotated songs per tag (left) and percentage of accumulated annotations of the first k most frequent tags (Magnatagatune).*

still very similar. The most frequently applied tag is *"rock"* and is set for 48.2% of all songs, while 569 out of 1072 tags are applied to less than 1%. Compared to the Magnatagatune dataset the number of tags in this dataset is far higher, while it does only contain about half the number of songs.

4.2 Evaluation Metrics

In our experiments all reported quality measures are first computed separately for each tag and are then averaged over all tags to come up with a global evaluation metric. For each tag we compute the following standard quality metrics: *f-Score*, *AUC-ROC* and the *accuracy*. Furthermore, we also report a modified variant of the *precision @k (p@k)* quality metric, which is called *precision @k above baseline (p@k AB)*. For each tag we first estimate the baseline precision at k, which is the expected precision when k samples are randomly drawn from the ground truth population without replacement. Obviously the baseline precision of each tag clearly depends on the individual class distribution. To reduce the influence of the individual class distribution of the tags on the overall metric, we first compute the $p@k$ and then subtract the estimated baseline, which then gives $p@k$ *AB*. For all our experiments we choose a fixed value of $k = 50$.

4.3 Results

We have evaluated both tag classification systems (*high* and *low*) on both tag classification datasets via a two fold-cross-validation. Table 1 shows the results. For both datasets we could identify a significant difference between the *high* and the *low* quality system. So we have ensured for both datasets that the learned

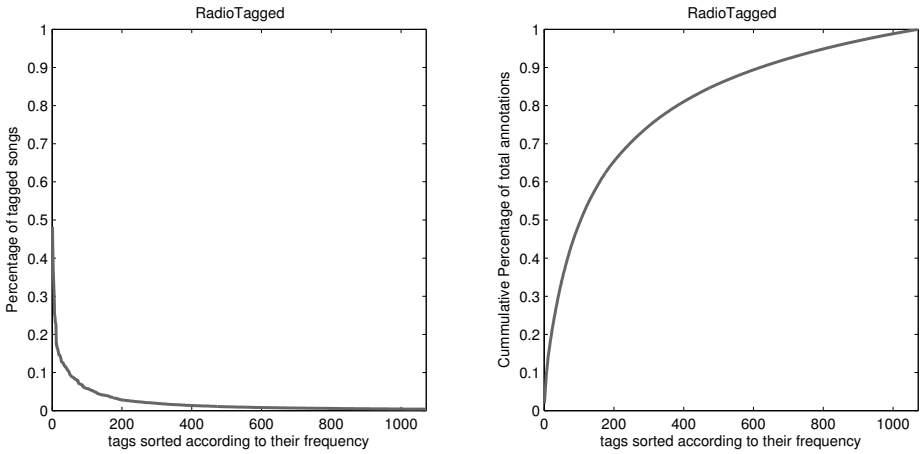

Fig. 3. *Percentage of annotated songs per tag (left) and percentage of accumulated annotations of the first k most frequent tags (RadioTagged).*

low and high quality auto-taggers differ significantly in terms of quality. In the next phase we will use these auto-taggers to predict tag profiles for songs of well-known music genre classification datasets to assess quality of a similarity measure based on these tag-predictors.

5 Phase 2: Estimate Tag Profile Similarities

In the previous phase we have ensured that the generated auto-taggers (*high* and *low*) differ significantly in terms of quality. Therefore, in case our assumption is correct we should end up with a higher quality music similarity measure for tag profiles estimated by the high quality tag classifiers compared to tag profiles estimated by the low quality tag classifiers. To actually estimate song similarities we follow the approach in [10] and compare the generated auto-tag profiles using the Manhattan distance.

Then, to assess the quality of both approaches (*low* and *high*) we evaluate them on six different genre classification datasets. In the following two subsections these datasets and the utilized quality measure are briefly introduced.

5.1 Datasets

To measure the quality of the resulting similarity estimates, we follow the standard approach in MIR and evaluate different approaches via nearest neighbour genre classification. In our experiments 6 different well-known genre classification datasets are used: *GTZAN*, *ISMIR 2004 Genre*, *ballroom*, *Homburg*, *Unique*, *1517-Artist*. It is worth mentioning that all these datasets are publicly available.

	Magnatagatune		**RadioTagged**	
	low	**high**	**low**	**high**
avg f-Score	0.1575	**0.2225***	0.0490	**0.0878***
AUC-ROC	0.6951	**0.8615***	0.5691	**0.6622***
Acc.	0.9719	**0.9749**	0.9608	**0.9640**
p@50 AB	0.1947	**0.2661***	0.0349	**0.0959***

Table 1. Qualitative comparison of the high quality (*high*) and the low quality (*low*) tag classification systems. Marked results (*) indicate statistically significant differences.

5.2 Evaluation Metrics

To assess the quality of the music similarity estimates of an algorithm the resulting similarity matrix containing all estimated pairwise distances among all songs in a collection is analyzed. The percentage of genre matches in the top k most similar songs is computed for each query song. To obtain an overall quality measure the per song results are averaged over the whole dataset. This quality measure is one of the automatic statistics that is computed at the MIREX Audio Music Similarity and Retrieval Task. There it is called *Genre Neighbourhood Clustering*, but is named *precision @k* (*p@k*) here. Interestingly, the results of the human music similarity evaluations at MIREX are year by year highly correlated with the *p@k* quality metric. Thus, this measure is an excellent choice to automatically assess the quality of music similarity systems. In our evaluation we will report the *precision @10* (p@10). For datasets *1517-Artist*, *Homburg* and *Unique* artist filtered results are reported.

5.3 Results

The results of this experiment are summarized in table 2. For both sets of classifiers, the one trained on the *Magnatagatune* dataset and the one trained on the *RadioTagged* dataset, the high quality auto-tag similarity measure outperforms the low quality version **on all six evaluation datasets**. Although this is inline with the intuitive expectations this result is a very encouraging, as the main implications of this experiment is that any improvement in automatic tag classification, will also lead to an improved content-based music similarity measure. Consequently, improvements both on the feature side and on the classification part can easily be integrated into an existing audio similarity algorithm.

6 Conclusions

In this paper we have focused on the relation between automatic tag classification and auto-tag based music similarity. Intuitively, auto-tag based music similarity algorithms should directly profit from qualitative improvements in automatic tag classification. Based on the results of our experiment we conclude

	Magnatagatune		**RadioTagged**	
Dataset	**low**	**high**	**low**	**high**
GTZAN	0.4569	**0.5765***	0.5459	**0.6167***
ISMIR 2004	0.7715	**0.7955**	0.7217	**0.7472**
ballroom	0.4089	**0.4645***	0.3983	**0.6143***
Homburg	0.3322	**0.4036***	0.3892	**0.4108***
Unique	0.5506	**0.6114***	0.5618	**0.6062***
1517-Artists	0.1839	**0.2355***	0.2128	**0.2475***

Table 2. Comparison of auto-tag based music similarity algorithms (*p@10*) based on *high* and *low* quality auto-taggers. Marked results (*) indicate statistically significant differences.

that this assumption is correct, which is a very encouraging result, as this defines a systematic and straightforward way to further improve content-based music similarity algorithms and content-based music recommender systems by improving the underlying automatic tag prediction systems.

Acknowledgements

This research is supported by the Austrian Science Funds (FWF): P22856-N23 and Z159.

References

1. A. Berenzweig, P. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *Proc. of the 2003 Int. Conf. on Multimedia and Expo (ICME-03)*, 2003.
2. T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.
3. D. Bogdanov, J. Serr, N. Wack, P. Herrera, and X. Serra. Unifying low-level and high-level music similarity measures. *IEEE Transaction on Multimedia*, 2010.
4. D. Eck, P. Lamere, T. B. Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Proc. of the 21st Conf. on Neural Infromation Processing Systems (NIPS-07)*, 2007.
5. B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, and P. Lambert. Content-based video description for automatic video genre categorization. In *of the 18th Int. Conf. on MultiMedia Modeling (MMM 2012)*, 2012.
6. E. Law and L. Ahn. Input-agreement: A new mechanism for collecting data using human computation games. In *Proc. of the 27th Int. Conf. on Human Factors in Computing Systems (CHI-09)*, 2009.
7. M. I. Mandel and D. P. W. Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2), 2008.
8. S. Ness, A. Theocharis, G. Tzanetakis, and L. Martins. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proc. of the 17th ACM Int. Conf. on Multimedia*, 2009.

9. K. Seyerlehner, M. Schedl, P. Knees, and R. Sonnleitner. A refined block-level feature set for classification, similarity and tag prediction. In *online Proc. of the 7th MIR Evaluation eXchange (MIREX-11)*, 2011.
10. K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees. Using block-level features for genre classification, tag classification and music similarity estimation. In *onl. Proc. of the 6th MIR Evaluation eXchange (MIREX-10)*, 2010.
11. K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees. Automatic music tag classification based on block-level features. In *Proc. of the 7th Sound and Music Computing Conference (SMC-10)*, 2010.
12. M. Slaney. Mixture of probability experts for audio retrieval and indexing. In *Proc. of the IEEE Int. Conf. on Multimedia and Expo*, 2002.
13. D. Turnbull, L. Barrington, and G. Lanckriet. Five approaches to collecting tags for music. In *Proc. of the 9th Int. Conf. on Music Information Retrieval*, 2008.
14. D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.
15. K. West and S. Cox. Incorporating cultural representations of features into audio music similarity estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3), 2010.
16. K. West, S. Cox, and P. Lamere. Incorporating machine-learning into music similarity estimation. In *Proc. of the 1st ACM Workshop on Audio and Music Computing Multimedia (AMCMM-06)*, 2006.
17. B. Whitman and R. Rifkin. Musical query-by-description as a multiclass learning problem. In *IEEE Multimedia Signal Processing Conf. (MMSP)*, 2002.