Learning to Tag from Open Vocabulary Labels

Edith Law, Burr Settles, and Tom Mitchell

Machine Learning Department
Carnegie Mellon University
{elaw,bsettles,tom.mitchell}@cs.cmu.edu

Abstract. Most approaches to classifying media content assume a fixed, closed vocabulary of labels. In contrast, we advocate machine learning approaches which take advantage of the millions of free-form tags obtainable via online crowd-sourcing platforms and social tagging websites. The use of such open vocabularies presents learning challenges due to typographical errors, synonymy, and a potentially unbounded set of tag labels. In this work, we present a new approach that organizes these noisy tags into well-behaved semantic classes using topic modeling, and learn to predict tags accurately using a mixture of topic classes. This method can utilize an arbitrary open vocabulary of tags, reduces training time by 94% compared to learning from these tags directly, and achieves comparable performance for classification and superior performance for retrieval. We also demonstrate that on open vocabulary tasks, human evaluations are essential for measuring the true performance of tag classifiers, which traditional evaluation methods will consistently underestimate. We focus on the domain of tagging music clips, and demonstrate our results using data collected with a human computation game called TagATune.

Keywords: Human Computation, Music Information Retrieval, Tagging Algorithms, Topic Modeling.

1 Introduction

Over the years, the Internet has become a vast repository of multimedia objects, organized in a rich and complex way through tagging activities. Consider music as a prime example of this phenomenon. Many applications have been developed to collect tags for music over the Web. For example, Last.fm is collaborative *social tagging* network which collects users' listening habits and roughly 2 million tags (e.g., "acoustic," "reggae," "sad," "violin") per month [12]. Consider also the proliferation of *human computation* systems, where people contribute tags as a by-product of doing a task they are naturally motivated to perform, such as playing causal web games. TagATune [14] is a prime example of this, collecting tags for music by asking two players to describe their given music clip to each other with tags, and then guess whether the music clips given to them are the same or different. Since deployment, TagATune has collected over a million annotations from tens of thousands of players.

In order to effectively organize and retrieve the ever-growing collection of music over the Web, many so-called *music taggers* have been developed [2,10,24] to automatically annotate music. Most previous work has assumed that the labels used to train music taggers come from a small fixed vocabulary and are devoid of errors, which greatly simplifies the learning task. In contrast, we advocate using tags collected by collaborative tagging websites and human computation games, since they leverage the effort and detailed domain knowledge of many enthusiastic individuals. However, such tags are noisy, i.e., they can be misspelled, overly specific, irrelevant to content (e.g., "albums I own"), and virtually unlimited in scope. This creates three main learning challenges: (1) *over-fragmentation*, since many of the enormous number of tags are synonymous or semantically equivalent, (2) *sparsity*, since most tags are only associated with a few examples, and (3) *scalability issues*, since it is computationally inefficient to train a classifier for each of thousands (or millions) of tags.

In this work, we present a new technique for classifying multimedia objects by tags that is scalable (i.e., makes full use of noisy, open-vocabulary labels that are freely available on the Web) and efficient (i.e., the training time remains reasonably short as the tag vocabulary grows). The main idea behind our approach is to organize these noisy tags into well-behaved semantic classes using a topic model [4], and learn to predict tags accurately using a mixture of topic classes. Using the TagATune [14] dataset as a case study, we compare the tags generated by our topic-based approach against a traditional baseline of predicting each tag independently with a binary classifier. These methods are evaluated in terms of both tag annotation and music retrieval performance. We also highlight a key limitation of traditional evaluation methods—comparing against a ground truth label set—which is especially severe for open-vocabulary tasks. Specifically, using the results from several Mechanical Turk studies, we show that human evaluations are essential for measuring the *true* performance of music taggers, which traditional evaluation methods will consistently underestimate.

2   Background

The ultimate goal of music tagging is to enable the automatic annotation of large collections of music, such that users can then browse, organize, and retrieve music in an semantic way. Although tag-based search querying is arguably one of the most intuitive methods for retrieving music, until very recently [2,10,24], most retrieval methods have focused on querying metadata such as artist or album title [28], similarity to an audio input query [6,7,8], or a small fixed set of category labels based on genre [26], mood [23], or instrument [9]. The lack of focus on music retrieval by rich and diverse semantic tags is partly due to a historical lack of labeled data for training music tagging systems.

A variety of machine learning methods have been applied to music classification, such as logistic regression [1], support vector machines [17,18], boosting [2], and other probabilistic models [10,24]. All of these approaches employ binary classifiers—one per label—to map audio features directly to a limited number

(tens to few hundreds) of tag labels independently. This is in contrast to the TagATune data set used in this paper, which has over 30,000 clips, over 10,000 unique tags collected from tens of thousands of users.

The drawback of learning to tag music from open-vocabulary training data is that it is *noisy*[1], by which we mean the over-fragmentation of the label space due to synonyms ("serene" vs. "mellow"), misspellings ("chello") and compound phrases ("guitar plucking"). Synonyms and misspellings cause music that belongs to the same class to be labeled differently, and compound phrases are often overly descriptive. All of these phenomena can lead to label sparsity, i.e., very few training examples for a given tag label.

It is possible to design data collection mechanisms to minimize such label noise in the first place. One obvious approach is to impose a controlled vocabulary, as in the Listen Game [25] which limits the set of tags to 159 labels pre-defined by experts. A second approach is to collect tags by allowing players to enter free-form text, but filter out the ones that have not been verified by multiple users, or that are associated with too few examples. For example, of the 73,000 tags acquired through the music tagging game MajorMiner [20], only 43 were used in the 2009 MIREX benchmark competition to train music taggers [15]. Similarly, the Magnatagatune data set [14] retains only tags that are associated with more than 2 annotators and 50 examples. Some recent work has attempted to mitigate these problems by distinguishing between content relevant and irrelevant tags [11], or by discovering higher-level concepts using tag co-occurrence statistics [13,16]. However, none of these works explore the use of these higher-level concepts in training music annotation or retrieval systems.

3   Problem Formulation

Assume we are given as training data a set of N music clips $\mathcal{C} = \{c_1, \ldots, c_N\}$ each of which has been annotated by humans using tags $\mathcal{T} = \{t_1, \ldots, t_V\}$ from a vocabulary of size V. Each music clip $c_i = (\boldsymbol{a_i}, \boldsymbol{x_i})$ is represented as a tuple, where $\boldsymbol{a_i} \in \mathbb{Z}^V$ is a the *ground truth tag* vector containing the frequency of each tag in $\mathcal{T}$ that has been used to annotate the music clip by humans, and $\boldsymbol{x_i} \in \mathbb{R}^M$ is a vector of M real-valued acoustic features, which describes the characteristics of the audio signal itself.

The goal of music annotation is to learn a function $\hat{f} : X \times T \to \mathbb{R}$, which maps the acoustic features of each music clip to a set of scores that indicate the relevance of each tag for that clip. Having learned this function, music clips can be retrieved for a search query q by rank ordering the distances between the query vector (which has value 1 at position j if the tag t_j is present in the search query, 0 otherwise) and the tag probability vector for each clip. Following [24], we measure these "distances" using KL divergence, which is a common information-theoretic measure of the difference between two distributions.

---

[1] We use *noise* to refer to the challenging side-effects of open tagging described here, which differs slightly from the common interpretation of mislabeled training data.
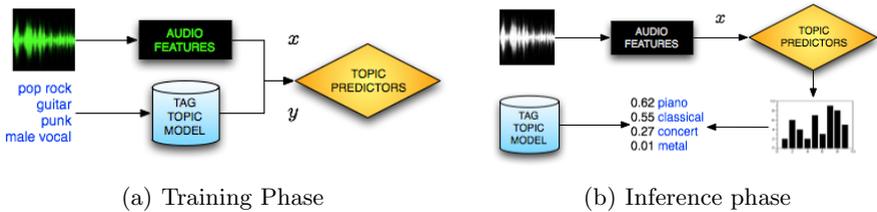
(a) Training Phase   (b) Inference phase

Fig. 1. The training and inference phases of the proposed approach

3.1   Topic Method (Proposed Approach)

We propose a new method for automatically tagging music clips, by first mapping from the music clip's audio features to a small number of semantic classes (which account for all tags in the vocabulary), and then generating output tags based on these classes. Training involves learning classes, or "topics," with their associated tag distributions, and the mapping from audio features to a topic class distribution. An overview of the approach is presented in Figure 1.

Training Phase. As depicted in Figure 1(a), training is a two-stage process. First, we induce a *topic model* [4,22] using the ground truth tags associated with each music clip in the training set. The topic model allows us to infer distribution over topics for each music clip in the training set, which we use to replace the tags as training labels. Second, we train a classifier that can predict topic class distributions directly from audio features.

In the first stage of training, we use Latent Dirichlet Allocation (LDA) [4], a common topic modeling approach. LDA is a hierarchical probabilistic model that describes a process for generating constituents of an entity (e.g., words of a document, musical notes in a score, or pixels in an image) from a set of latent class variables called topics. In our case, constituents are tags and an entity is the semantic description of a music clip (i.e., set of tags). Figure 2(a) shows an example model of 10 topics induced from music annotations collected by TagATune. Figure 2(b) and Figure 2(c) show the topic distributions for two very distinct music clips and their ground truth annotations (in the caption; note synonyms and typos among the tags entered by users). The music clip from Figure 2(b) is associated with both topic 4 (classical violin) and topic 10 (female opera singer). The music clip from Figure 2(c) is associated with both topic 7 (flute) and topic 8 (quiet ambient music).

In the second stage of training, we learn a function that maps the audio features for a given music clip to its topic distribution. For this we use a maximum entropy (MaxEnt) classifier [5], which is a multinomial generalization of logistic regression. We use the LDA and MaxEnt implementations in the MALLET toolkit[2], with a slight modification of the optimization procedure [29] which enables us to train a MaxEnt model from class distributions rather than a single class label. We refer to this as the *Topic Method*.

---

[2] http://mallet.cs.umass.edu

1	electronic beat fast drums synth dance beats jazz
2	male choir man vocal male_vocal vocals choral singing
3	indian drums sitar eastern drum tribal oriental middle_eastern
4	classical violin strings cello violins classic slow orchestra
5	guitar slow strings classical country harp solo soft
6	classical harpsichord fast solo strings harpsicord classic harp
7	flute classical flutes slow oboe classic clarinet wind
8	ambient slow quiet synth new_age soft electronic weird
9	rock guitar loud metal drums hard_rock male fast
10	opera female woman vocal female_vocal singing female_voice vocals

(a) Topic Model
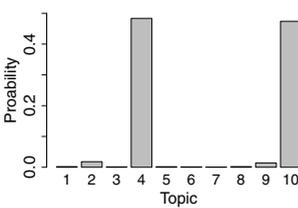
(b) woman, classical, classsical, opera, male, violen, violin, voice, singing, strings, italian
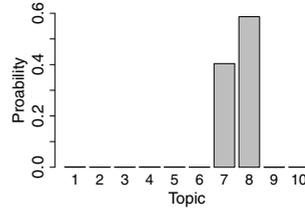
(c) chimes, new age, spooky, flute, quiet, whistle, fluety, ambient, soft, high pitch, bells

Fig. 2. An example LDA model of 10 topic classes learned over music tags, and the representation of two sample music clips annotations by topic distribution

Our approach tells an interesting generative story about how players of TagATune might decide on tags for the music they are listening to. According to the model, each listener has a latent topic structure in mind when thinking of how to describe the music. Given a music clip, the player first selects a topic according to the topic distribution for that clip (as determined by audio features), and then selects a tag according to the posterior distribution of the chosen topics. Under this interpretation, our goal in learning a topic model over tags is to discover the topic structure that the players use to generate tags for music, so that we can leverage a similar topic structure to automatically tag new music.

Inference Phase. Figure 1(b) depicts the process of generating tags for novel music clips. Given the audio features $\boldsymbol{x_i}$ for a test clip c_i, the trained MaxEnt classifier is used to predict a topic distribution for that clip. Based on this predicted topic distribution, each tag t_j is then given a relevance score $P(t_j|\boldsymbol{x_i})$ which is its expected probability over all topics:

$$P(t_j|\boldsymbol{x_i}) = \sum_{k=1}^{K} P(t_j|y_k)P(y_k|\boldsymbol{x_i}),$$

where $j = 1, \ldots, V$ ranges over the tag vocabulary, and $k = 1, \ldots, K$ ranges over all topic classes in the model.

3.2   Tag Method (Baseline)

To evaluate the efficiency and accuracy of our method, we compare it against an approach that predicts $P(t_j|\boldsymbol{x_i})$ directly using a set of binary logistic regression classifiers (one per tag). This second approach is consistent with previous approaches to music tagging with closed vocabularies [1,17,18,2,10,24]. We refer to it as the *Tag Method*. In some experiments we also compare against a method that assigns tags randomly.
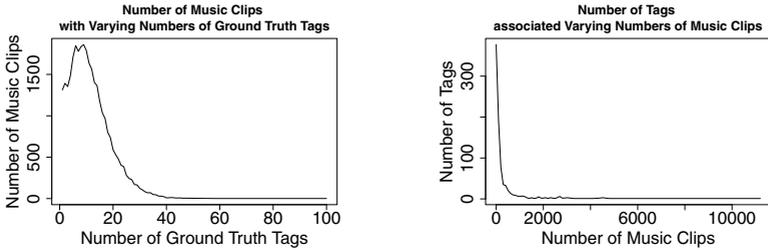
4   Data Set

The data is collected via a two-player online game called called TagATune [14]. Figure 3 shows the interface of TagATune. In this game, two players are given either the same or different music clips, and are asked to describe their given music clip. Upon reviewing each other's description, they must guess if the music clips are the same or different.

There exist several human computation games [20,25] that collect tags for music that are based on the *output-agreement mechanism* (a.k.a. the ESP Game [27] mechanism), where two players must match on a tag in order for that tag to become a valid label for a music clip. In our previous work [14], we have showed that output-agreement games, although effective for image annotation, are restrictive for music data: there are so many ways to describe music and sounds that players often have a difficult time agreeing on any tags. In TagATune, the problem of agreement is alleviated by allowing players to communicate with each other. Furthermore, by requiring that the players guess whether the music are the same or different based on each other's tags, the quality and validity of the tags are ensured. The downside of opening up the communication between players is that the tags entered are more noisy.

Fig. 3. A screen shot of the TagATune user interface

(a) Number of music clips with X number of ground truth tags

(b) Number of tags associated X number of music clips

Fig. 4. Characteristics of the TagATune data set

Figure 4 shows the characteristics of the TagATune dataset. Figure 4(a) is a rank frequency plot showing the number of music clips that have a certain number of ground truth tags. The plot reveals a disparity in the number of ground truth tags each music clip has – a majority of the clips (1,500+) have under 10, approximately 1,300 music clips have only 1 or 2, and very few have a large set (100+). This creates a problem in our evaluation – many of the generated tags that are relevant for the clip may be missing from the ground truth tags, and therefore will be considered incorrect. Figure 4(b) is a rank frequency plot showing the number of tags that have a certain number of music clips available to them as training examples. The plot shows that the vast majority of the tags have few music clips to use as training examples, while a small number of tags are endowed with a large number of examples. This highlights the aforementioned sparsity problem that emerges when tags are used directly as labels, a problem that is addressed by our proposed method.

We did a small amount of pre-processing on a subset of the data set, tokenizing tags, removing punctuation and four extremely common tags that are not related to the content of the music, i.e. "yes," "no," "same," "diff". In order to accommodate the baseline Tag Method, which requires a sufficient number of training examples for each binary classification task, we also eliminated tags that have fewer than 20 training music clips. This reduces the number of music clips from 31,867 to 31,251, the total number of ground truth tags from 949,138 to 699,440, and the number of *unique* ground truth tags from 14,506 to 854. Note that we are throwing away a substantial amount of tag data in order to accommodate the baseline Tag Method. A key motivation for using our Topic Method is that we do not need to throw away any tags at all. Rare tags, i.e. tags that are associated with only one or two music clips, can still be grouped into a topic, and used in the annotation and retrieval process.

Each of the 31,251 music clips is 29 seconds in duration, and is represented by a set of ground truth tags collected via the TagATune game, as well as a set of content-based (spectral and temporal) audio features extracted using the technique described in [19].

5   Experiments

We conducted several experiments guided by five central questions about our
proposed approach. (1) *Feasibility*: given a set of noisy music tags, is it possible to
learn a low-dimensional representation of the tag space that is both semantically
meaningful and predictable by music features? (2) *Efficiency*: how does training
time compare against the baseline method? (3) *Annotation Performance*: how
accurate are the generated tags? (4) *Retrieval Performance*: how well do the
generated tags facilitate music retrieval? (5) *Human Evaluation*: to what extent
are the performance evaluations a reflection of the true performance of the music
taggers? All results are averaged over five folds using cross-validation.

5.1   Feasibility

Table 1 (on the next page) shows the top 10 words for each topic learned by LDA
with the number of topics fixed at 10, 20 and 30. In general, the topics are able to
capture meaningful groupings of tags, e.g., synonyms (e.g., "choir/choral/cho-
rus" or "male/man/male_vocal"), misspellings (e.g., "harpsichord/harpsicord"
or "cello/chello"), and associations (e.g., "indian/drums/sitar/eastern/oriental"
or "rock/guitar/loud/metal"). As we increase the number of topics, new seman-
tic grouping appear that were not captured by models which use a fewer number
of topics. For example, in 20-topic model, topic 3 (which describes soft classical
music), topic 13 (which describes jazz), and topic 17 (which describes rap, hip-
hop and reggae) are new topics that are not evident in the model with only 10
topics. We also observe some repetition or refinement of topics as the number
of topic increases (e.g., topics 8, 25 and 27 in the 30-topic model all describe
slightly different variations on female vocal music).

It was difficult to know exactly how many topics can succinctly capture the
concepts underlying the music in our data set. Therefore, in all our experiments
we empirically tested how well the topic distribution and the best topic can be
predicted using audio features, fixing the number of topics at 10, 20, 30, 40, and
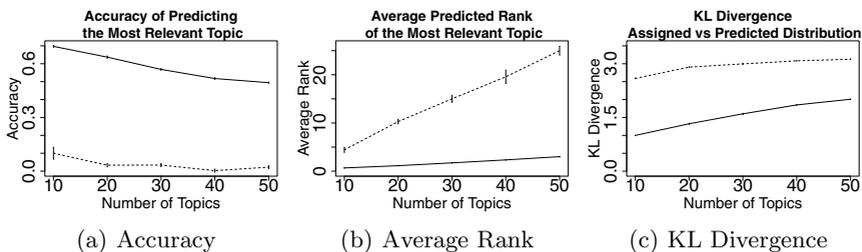50. Figure 5 summarizes the results. We evaluated performance using several

(a) Accuracy   (b) Average Rank   (c) KL Divergence

Fig. 5. Results showing how well topic distributions or the best topic can be predicted
from audio features. The metrics include accuracy and average rank of the most relevant
topic, and KL divergence between the assigned and predicted topic distribution.

Table 1. Topic Model with 10, 20, and 30 topics. The topics in bold in the 20-topic model are examples of new topics that emerge when the number of topics is increased from 10 to 20. The topics marked by * in the 30-topic model are examples of topics that start to repeat as the number of topics is increased.

	10 Topics
1	electronic beat fast drums synth dance beats jazz electro modern
2	male choir man vocal male_vocal vocals choral singing male_voice pop
3	indian drums sitar eastern drum tribal oriental middle_eastern foreign fast
4	classical violin strings cello violins classic slow orchestra string solo
5	guitar slow strings classical country harp solo soft quiet acoustic
6	classical harpsichord fast solo strings harpsicord classic harp baroque organ
7	flute classical flutes slow oboe classic clarinet wind pipe soft
8	ambient slow quiet synth new_age soft electronic weird dark low
9	rock guitar loud metal drums hard_rock male fast heavy male_vocal
10	opera female woman vocal female_vocal singing female_voice vocals female_vocals voice

	20 Topics
1	indian sitar eastern oriental strings middle_eastern foreign guitar arabic india
2	flute classical flutes oboe slow classic pipe wind woodwind horn
3	**slow quiet soft classical solo silence low calm silent very_quiet**
4	male male_vocal man vocal male_voice pop vocals singing male_vocals guitar
5	cello violin classical strings solo slow classic string violins viola
6	opera female woman vocal classical vocal singing female_opera female_vocal female_voice operatic
7	female woman vocal female_vocal singing female_voice vocals female_vocals pop voice
8	guitar country blues folk irish banjo fiddle celtic harmonica fast
9	guitar slow classical strings harp solo classical_guitar soft acoustic spanish
10	electronic synth beat electro ambient weird new_age drums electric slow
11	drums drum beat beats tribal percussion indian fast jungle bongos
12	fast beat electronic dance beats synth electro trance loud
13	**jazz jazzy drums sax bass funky guitar funk trumpet clapping**
14	ambient slow synth new_age electronic weird quiet soft dark drone
15	classical violin strings violins classic orchestra slow string fast cello
16	harpsichord classical harpsicord strings baroque harp classic fast medieval harps
17	**rap talking hip_hop voice reggae male male_voice man speaking voices**
18	classical fast solo organ classic slow soft quick upbeat light
19	choir choral opera chant chorus vocal vocals singing voices chanting
20	**rock guitar loud metal hard_rock drums fast heavy electric_guitar heavy_metal**

	30 Topics
1	choir choral opera chant chorus vocal male chanting vocals singing
2	classical solo classic oboe fast slow clarinet horns soft flute
3	rap organ talking hip_hop voice speaking man male_voice male man_talking
4	rock metal loud guitar hard_rock heavy fast heavy_metal male punk
5	guitar classical slow strings solo classical_guitar acoustic soft harp spanish
6	cello violin classical strings solo slow classic string violins chello
7	violin classical strings violins classic slow cello string orchestra baroque
*8	female woman female_vocal vocal female_voice pop singing female_vocals vocals voice
9	bells chimes bell whistling xylophone whistle chime weird high_pitch gong
10	ambient slow synth new_age electronic soft spacey instrumental quiet airy
11	rock guitar drums loud electric_guitar fast pop guitars electric bass
12	slow soft quiet solo classical sad calm mellow very_slow low
13	water birds ambient rain nature ocean waves new_age wind slow
14	irish violin fiddle celtic folk strings clapping medieval country violins
15	electronic synth beat electro weird electric drums ambient modern fast
16	indian sitar eastern middle_eastern oriental strings arabic guitar india foreign
17	drums drum beat beats tribal percussion indian fast jungle bongos
18	classical strings violin orchestra violins classic orchestral string baroque fast
19	quiet slow soft classical silence low very_quiet silent calm solo
20	flute classical flutes slow wind woodwind classic soft wind_instrument violin
21	guitar country banjo folk harmonica bluegrass acoustic twangy fast
22	male man male_vocal vocal male_voice pop singing vocals male_vocals voice
23	jazz jazzy drums sax funky funk bass guitar trumpet reggae
24	harp strings guitar dulcimer classical sitar slow string oriental plucking
*25	vocal vocals singing foreign female voices women woman voice choir
26	fast loud upbeat quick fast_paced very_fast happy fast_tempo fast_beat faster
*27	opera female woman vocal classical singing female_opera female_voice female_vocal operatic
28	ambient slow dark weird drone low quiet synth electronic eerie
29	harpsichord classical harpsicord baroque strings classic harp medieval harps guitar
30	beat fast electronic dance drums beats synth electro trance upbeat

metrics, including accuracy and average rank of the most probable topic, as well as the KL divergence between the ground truth topic distribution and the predicted distribution.

Although we see a slight degradation of performance as the number of topics increases, all models significantly outperform the random baseline, which uses random distributions as labels for training. Moreover, even with 50 topics, the average rank of the top topic is still around 3, which suggests that the classifier is capable of predicting the most relevant topic, an important pre-requisite for the generation of accurate tags.

5.2   Efficiency

A second hypothesis is that the Topic Method would be more computation-ally efficient to train, since it learns to predict a joint topic distribution in a reduced-dimensionality tag space (rather than a potentially limitless number of independent classifiers). Training the Topic Method (i.e., inducing the topic model *and* the training the classifier for mapping audio features to a topic distri-bution) took anywhere from 18.3 minutes (10 topics) to 48 minutes (50 topics) per fold, but quickly plateaus after 30 topics: ▄▅▇▇ . The baseline Tag Method, by contrast, took 845.5 minutes (over 14 hours) per fold. Thus, the topic ap-proach can reduce training time by 94% compared to the Tag Method baseline, which confirms our belief that the proposed method will be significantly more scalable as the size of the tag vocabulary grows, while eliminating the need to filter low-frequency tags.

5.3   Annotation Performance

Following [10], we evaluate the accuracy of the 10 tags with the highest proba-bilities for each music clip, using three different metrics: per-clip metric, per-tag metric, and omission-penalizing per-tag metric.

Per-Clip Metrics. The per-clip precision@N metric measures the proportion of correct tags (according to agreement with the ground truth set) amongst the N most probable tags for each clip according to the tagger, averaged over all the clips in the test set. The results are presented in Figure 6. The Topic Method and baseline Tag Method both significantly outperform the random baseline, and the Topic Method with 50 topics is indistinguishable from the Tag Method.

Per-Tag Metric. Alternatively, we can evaluate the annotation performance by computing the precision, recall, and F-1 scores for each tag, averaged over all the tags that are output by the algorithm (i.e. if the music tagger does not output a tag, it is ignored). Specifically, given a tag t, we calculate its precision $P_t = \frac{c_t}{a_t}$, recall $R_t = \frac{c_t}{g_t}$, and and F-1 measure $F_t = \frac{2 \times P_t \times R_t}{P_t + R_t}$, where g_t is the number of test music clips that have t in their ground truth sets, a_t is the number of clips that are annotated with t by the tagger, and c_t is the number of clips that have been *correctly* annotated with the tag t by the tagger (i.e., t is found in the ground truth set). The overall per-tag precision, recall and F-1
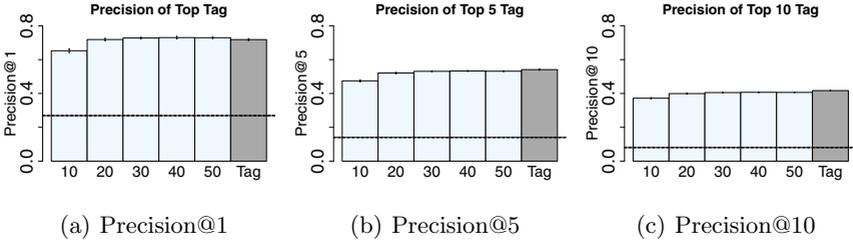
Fig. 6. Per-clip Metrics. The light-colored bars represent Topic Method with 10, 20, 30, 40 and 50 topics. The dark-colored bar represents the Tag Method. The horizontal line represent the random baseline, and the dotted lines represent its standard deviation.
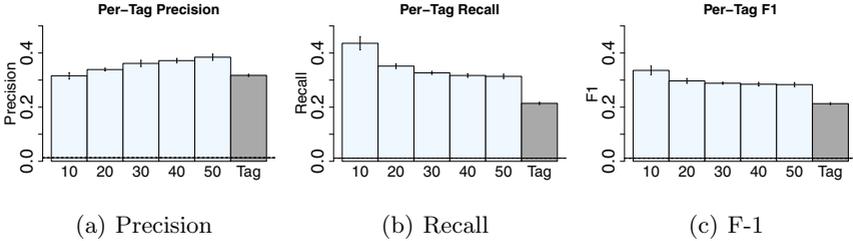

Fig. 7. Per-tag Metrics. The light-colored bars represent Topic Method with 10, 20, 30, 40 and 50 topics. The dark-colored bar represents the Tag Method. The horizontal line represent the random baseline, and the dotted lines represent its standard deviation.

scores for a test set are P_t, R_t and F_t for each tag t, averaged over all tags in the vocabulary. Figure 7 presents these results, showing that the Topic Method significantly outperforms the baseline Tag Method under this set of metrics.

Omission-Penalizing Per-Tag Metrics. A criticism of some of the previous metrics, in particular the *per-clip* and *per-tag precision* metrics, is that a tagger that simply outputs the most common tags (omitting rare ones) can still perform reasonably well. Some previous work [2,10,24] has adopted a set of per-tag metrics that penalize omissions of tags that could have been used to annotate music clips in the test set. Following [10,24], we alter tag precision P_t to be the empirical frequency E_t of the tag t in the test set *if* the tagger failed to predict t for any instances at all (otherwise, $P_t = \frac{c_t}{a_t}$ as before). Similarly, the tag recall $R_t = 0$ if the tagger failed to predict t for any music clips (and $R_t = \frac{c_t}{g_t}$ otherwise). This specification penalizes classifiers that leave out tags, especially rare ones. Note these metrics are upper-bounded by a quantity that depends on the number of tags output by the algorithm. This quantity can be computed empirically by setting the precision and recall to 1 when a tag is present, and to E_t and 0 (respectively) when a tag is omitted.

Results (Figure 8) show that for the Topic Method, performance increases with more topics, but reaches a plateau as the number of topics approaches 50. One possible explanation is revealed by Figure 9(a), which shows that the number
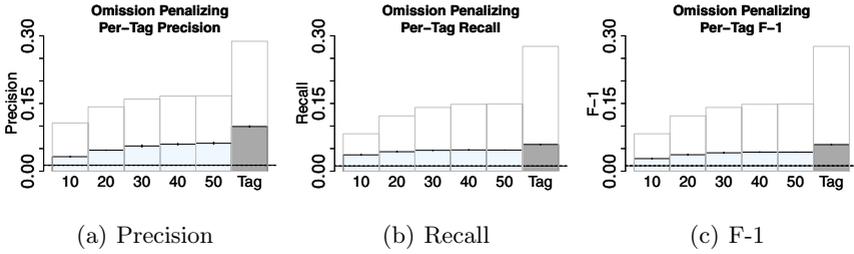
Fig. 8. Omission-Penalizing Per-tag Metrics. Light-colored bars represent the Topic Method with 10, 20, 30, 40 and 50 topics. Dark-colored bars represent the Tag Method. Horizontal lines represent the random baseline. Grey outlines indicate upper bounds.
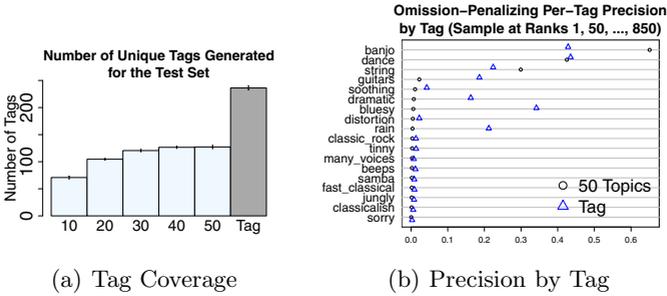

Fig. 9. Tag coverage and loss of precision due to omissions

of unique tags generated by the Topic Method reaches a plateau at around this point. In additional experiments using 60 to 100 topics, we found that this plateau persists. This might explain why the Tag Method outperforms the Topic Method under this metric—it generates many more unique tags.

Figure 9(b), which shows precision scores for sample tags achieved by each method, confirms this hypothesis. For the most common tags (e.g., "banjo," "dance," "string"), the Topic Method achieves superior or comparable precision, while for rarer tags (e.g., "dramatic," "rain" etc.), the Tag Method is better and the Topic Method receives lower scores due to omissions. Note that these low-frequency tags contain more noise (e.g., "jungly," "sorry"), so it could be that the Tag Method is superior simply on its ability to output noisy tags.

5.4   Retrieval Performance

The tags generated by a music tagger can be used to facilitate retrieval. Given a search query, music clips can be ranked by the KL divergence between the query tag distribution and the tag probability distribution for each clip. We measure the quality of the top 10 music clips retrieved using the mean average precision [24] metric, $M_{10} = \frac{1}{10} \sum_{r=1}^{10} \frac{s_r}{r}$, where s_r is the number of "relevant" (i.e., the search query can be found in the ground truth set) songs at rank r.
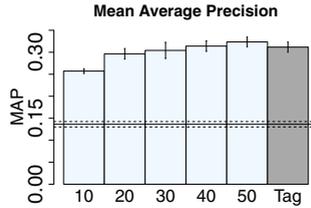
Mean Average Precision

Fig. 10. Retrieval performance in terms of mean average precision

Figure 10 shows the performance of the three methods under this metric. The retrieval performance of the Topic Method with 50 topics is slightly better than the Tag Method, but otherwise indistinguishable. Both methods perform significantly better than random (the horizontal line).

5.5   Human Evaluation

We argue that the performance metrics used so far can only *approximate* the quality of the generated tags. The reason is that generated tags that cannot be found amongst ground truth tags (due to missing tags or vocabulary mismatch) are counted as wrong, when they might in fact be relevant but missing due to the subtleties of using an open tag vocabulary.

In order to compare the true merit of the tag classifiers, we conducted several Mechanical Turk experiments asking humans to evaluate the annotation and retrieval capabilities of the Topic Method (with 50 topics), Tag Method and Random Method. For the annotation task, we randomly selected a set of 100 music clips, and solicited evaluations from 10 unique evaluators per music clip. For each clip, the user is given three lists of tags generated by each of the three methods. The order of the lists is randomized each time to eliminate presentation bias. The users are asked to (1) click the checkbox beside a tag if it describes the music clip well, and (2) rank order their overall preference for each list.

Figure 11 shows the per-tag precision, recall and F-1 scores as well as the per-clip precision scores for the three methods, using both ground truth set evaluation and using human evaluators. Results show that when tags are judged based on whether they are present in the ground truth set, performance of the tagger is grossly underestimated for all metrics. In fact, of the predicted tags that the users considered "appropriate" for a music clip (generated by either the Topic Method or the Tag Method method), on average, approximately half of them are missing from the ground truth set.

While the human-evaluated performance of the Tag Method and Topic Method are virtually identical, when asked to rank the tag lists evaluators preferred the the Tag Method (62.0% of votes) over the Topic Method (33.4%) or Random (4.6%). Our hypothesis is that people prefer the Tag Method because its has better coverage (Section 5.3). Since evaluation is based on 10 tags generated by the tagger, we conjecture that a new way of generating this *set* of
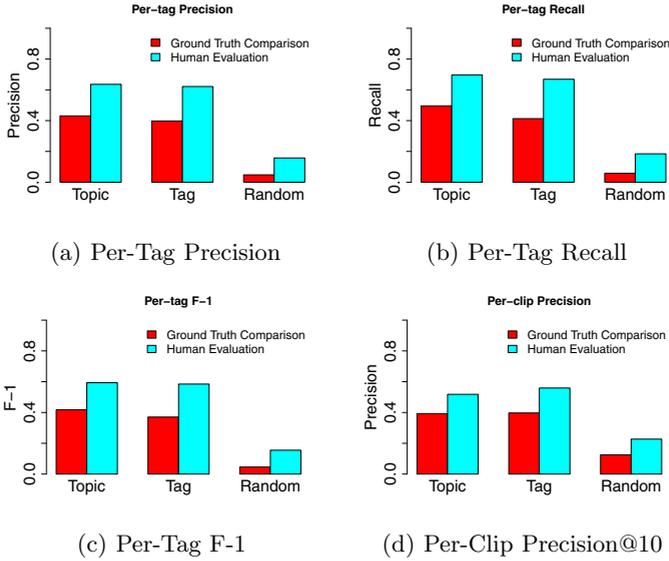
(a) Per-Tag Precision

(b) Per-Tag Recall

(c) Per-Tag F-1

(d) Per-Clip Precision@10

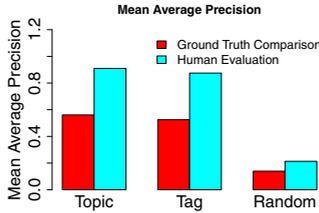**Fig. 11.** Mechanical Turk results for annotation performance

Fig. 12. Mechanical Turk results for music retrieval performance

output tags from topic posteriors (e.g., to improve diversity) may improve in this regard.

We also conducted an experiment to evaluating retrieval performance, where we provided each human evaluator a single-word search query and three lists of music clips retrieved by each method. We used 100 queries and 3 evaluators per query. Users were asked to check each music clip that they considered to be "relevant" for the query. In addition, they are asked to rank order the three lists in terms of their overall relevance to the query.

Figure 12 shows the mean average precision, when the ground truth tags versus human judgment is used to evaluate the relevance of each music clip in the retrieved set. As with annotation performance, the performance of all methods is significantly lower when evaluated using the ground truth set than when using human evaluations. Finally, when asked to rank music lists, users strongly preferred our Topic Method (59.3% of votes) over the Tag Method (39.0%) or Random (1.7%).

6   Conclusion and Future Work

The purpose of this work is to show how tagging algorithms can be trained, in an efficient way, to generate labels for objects (e.g., music clips) when the training data consists of a huge vocabulary of noisy labels. Focusing on music tagging as the domain of interest, we showed that our proposed method is both time and data efficient, while capable of achieving comparable (or superior, in the case of retrieval) performance to the traditional method of using tags as labels directly. This work opens up the opportunity to leverage the huge number of tags freely available on the Web for training annotation and retrieval systems.

Our work also exposes the problem of evaluating tags when the ground truth sets are noisy or incomplete. Following the lines of [15], an interesting direction would be to build a human computation game that is suited specifically for evaluating tags, and which can become a service for evaluating any music tagger.

There have been recent advances on topic modeling [3,21] that induce topics not only text, but also from other metadata (e.g., audio features in our setting). These methods may be good alternatives for training the topic distribution classifier in a one-step process as opposed to two, although our preliminary work in this direction has so far yielded mixed results.

Finally, another potential domain for our Topic Method is birdsong classification. To date, there are not many (if any) databases that allow a birdsong search by arbitrary tags. Given the myriad ways of describing birdsongs, it would be difficult to train a tagger that maps from audio features to tags directly, as most tags are likely to be associated with only a few examples. In collaboration with Cornell's Lab of Ornithology, we plan to use TagATune to collect birdsong tags from the tens of thousands of "citizen scientists" and apply our techniques to train an effective birdsong tagger and semantic search engine.

References

1. Bergstra, J., Lacoste, A., Eck, D.: Predicting genre labels for artists using freedb. In: ISMIR, pp. 85–88 (2006)
2. Bertin-Mahieux, T., Eck, D., Maillet, F., Lamere, P.: Autotagger: a model for predicting social tags from acoustic features on large music databases. TASLP 37(2), 115–135 (2008)
3. Blei, D., McAuliffe, J.D.: Supervised topic models. In: NIPS (2007)
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
5. Csiszar, I.: Maxent, mathematics, and information theory. In: Hanson, K., Silver, R. (eds.) Maximum Entropy and Bayesian Methods. Kluwer Academic Publishers, Dordrecht (1996)
6. Dannenberg, R.B., Hu, N.: Understanding search performance in query-by-humming systems. In: ISMIR, pp. 41–50 (2004)

7. Eisenberg, G., Batke, J.M., Sikora, T.: Beatbank – an mpeg-7 compliant query by tapping system. Audio Engineering Society Convention, 6136 (2004)
8. Goto, M., Hirata, K.: Recent studies on music information processing. Acoustic Science and Technology, 419–425 (2004)
9. Herrera, P., Peeters, G., Dubnov, S.: Automatic classification of music instrument sounds. Journal of New Music Research, 3–21 (2003)
10. Hoffman, M., Blei, D., Cook, P.: Easy as CBA: A simple probabilistic model for tagging music. In: ISMIR, pp. 369–374 (2009)
11. Iwata, T., Yamada, T., Ueda, N.: Modeling social annotation data with content relevance using a topic model. In: NIPS (2009)
12. Lamere, P.: Social tagging and music information retrieval. Journal of New Music Research 37(2), 101–114 (2008)
13. Laurier, C., Sordo, M., Serra, J., Herrera, P.: Music mood representations from social tags. In: ISMIR, pp. 381–386 (2009)
14. Law, E., von Ahn, L.: Input-agreement: A new mechanism for collecting data using human computation games. In: CHI, pp. 1197–1206 (2009)
15. Law, E., West, K., Mandel, M., Bay, M., Downie, S.: Evaluation of algorithms using games: The case of music tagging. In: ISMIR, pp. 387–392 (2009)
16. Levy, M., Sandler, M.: A semantic space for music derived from social tags. In: ISMIR (2007)
17. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: SIGIR, pp. 282–289 (2003)
18. Mandel, M., Ellis, D.: Song-level features and support vector machines for music classification. In: ISMIR (2005)
19. Mandel, M., Ellis, D.: Labrosa's audio classification submissions (2009)
20. Mandel, M., Ellis, D.: A web-based game for collecting music metadata. Journal of New Music Research 37(2), 151–165 (2009)
21. Mimno, D., McCallum, A.: Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In: UAI (2008)
22. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Erlbaum, Hillsdale (2007)
23. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music emotions. In: ISMIR, pp. 325–330 (2008)
24. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. TASLP 16(2), 467–476 (2008)
25. Turnbull, D., Liu, R., Barrington, L., Lanckriet, G.: A game-based approach for collecting semantic annotations of music. In: ISMIR, pp. 535–538 (2007)
26. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10(5), 293–302 (2002)
27. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: CHI, pp. 319–326 (2004)
28. Whitman, B., Smaragdis, P.: Combining musical and cultural features for intelligent style detection. In: ISMIR (2002)
29. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: KDD, pp. 937–946 (2009)