IMPROVING MUSIC AUTO-TAGGING WITH TRIGGER-BASED CONTEXT MODEL

Qin Yan, Cong Ding, Jingjing Yin, Yong Lv

College of Computer and Information, Hohai University, Nanjing, China
yan_qin@hhu.edu.cn, monkey.dcong1@gmail.com, jingjing.yin91@gmail.com, lynetwork@gmail.com

ABSTRACT

Music auto-tagging has been an active research topic as it learns the relationship between the content of audio tracks and semantic tags such that users can query by both tags and audio segments without being troubled by the cold start problem. In this paper, we propose a new trigger-based context model to refine the existing content model based auto-tagging systems. The trigger based context model improves accruacy of weakly labeled tags in "Genre", "Solo" and "Usage" by 10.63%, 10% and 26.43% respectively, which are usually poorly modeled due to lack of data in the content model based systems. Experiment results indicate that a combination of the content and context models outperforms the content based only auto-tagging system and the baseline Turnbull's MixHier model by 0.74% and 2.64% in average precision rate respectively.

Index Terms—Music auto-tagging improvement, context model, trigger feature selection, maximum entropy

1. INTRODUCTION

With the continuous increasing scale of music volume and the well-known cold start problem (tracks without manual annotations are inaccessible), music auto-tagging has become a popular topic in MIR research to bridge the semantic gap between low-level computable audio features and high-level human perceptible labels [1]. Music auto-tagging is defined as a multi-label classification task which aims to describe the semantic content of given tracks by assigning related tags to them automatically. Considering diverse musical tags including emotion, genre, instrument and vocals, well-trained music auto-tagger can be used towards not only offline music collections indexing but also online music discovery.

A typical auto-tagging system employs a content-based method, which includes two parts: a set of acoustic features extracted from audio signals and statistic models trained independently for different tags. In such a system, manual annotated training data is learnt by statistic models so that audio feature space can be transformed into semantic space (normally with hundreds of tags in it). Then each test track is automatically labeled with a subset of best suitable tags by comparing extracted audio features to corresponding tag models. The output probabilities of testing tracks are called semantic multinomials (SMNs) [2].

It is well known that the quality and quantity of the training data greatly affects the performance of auto-tagging system. It is evident that the more the diversified tags exist, the higher possibilities that some of them are under modelled due to limited presence in training sets [3]. These so-called "weakly labeled tags" suffer a lot from the imbalance problem especially in systems where tags are modelled independently. Although learning from imbalance data has been deeply studied for binary and multi-classification tasks (using approaches such as data preprocessing and cost sensitive classification), there are little approaches dealing with such imbalance problem in music auto-tagging [4-6].

Variable feature sets and tag-level Gaussian mixture models are used to construct the basic content based auto-tagging system in our previous work [7]. In this paper we endeavor to solve the problem of "weakly labeled tags" by improving the existing auto-tagging systems in two aspects: the semantic context analysis and the trigger-based context modeling. The context analysis captures aspects beyond pure content from audio signals and different levels of correlations between tags are then apprehended by trigger-based context models with maximum entropy approach. The maximum entropy approach has been widely used in natural language processing (NLP), including part of speech tagging, parse selection, sentence boundary detection and ambiguity resolution [8-9]. It has also been applied to other fields as computer vision [10] and document annotation in the biomedical domain [11] and performs with high accuracy in recent years.

The rest of the paper is organized as below. Section 2 gives an overview of context based auto-tagging refinement. Tag correlation and trigger features are discussed in Section 3. Section 4 presents the trigger-based context models and the probability refinement criterion followed by the performance evaluation and comparison in Section 5.

2. SYSTEM OVERVIEW

The structure of the proposed two-stage music auto-tagging system is illustrated in Fig.1. In stage 1, the content model is obtained by training Gaussian mixture models (GMMs) with

This work was supported in part by National Natural Science Foundation of China (No.61170297, 61301218) and Qing Lan Project of Jiangsu Province of China.

variable feature sets while in stage 2 tag correlations are identified to establish the context model for each tag via maximum entropy (ME) approach. In a word stage 2 essentially refines the imperfect output tag probabilities of the Stage 1. The second stage is the focus of this paper while the work in the first stage has been reported in [7]. To facilitate the quantitative evaluation, the experiments are conducted on Computer Audio Lab 500 (CAL500) [2].

3. TRIGGER FEATURES

To extract the correlation information from the manually labeled semantic tags of the training dataset, we borrow the idea of trigger pair from the language processing field [12] as the information bearing element. This section briefs the measure of tag correlation and trigger features selection before constructing the contextual model.

3.1. Tags Correlation

There are several similarity measures available for extracting tag correlations such as matching similarity, overlap similarity, cosine similarity, Jaccard similarity, Dice coefficient and mutual information [13]. In this work the projected Dice coefficient is employed due to its simplicity. $W = \{w_i\},\ i = 1,2, \dots, |W|$ is the tag vocabulary, where each tag w_i represents a class and $|W|$ is total number of tags. If tag w_i is strongly correlated with another tag $w_k \in W$, then $(w_k \to w_i)$ is considered as a trigger pair. w_k is the trigger tag and w_i is the triggered tag. If tag w_k is annotated to a track, it will trigger tag w_i and raise the probability w_i given the track. The correlation of trigger pair $(w_k \to w_i)$ is computed as

$$C_{ik} = \frac{2N_{w_i \cap w_k}}{N_{w_i} + N_{w_k}} \tag{1}$$

where N_{w_i} represents the number of tracks positive associated with tag w_i, and $N_{w_i \cap w_k}$ is the number of tracks annotated by both tag w_i and w_k. The correlation value C_{ik} ranges between 0 and 1. $C_{ik} > 0$ if there is a positive correlation. The higher C_{ik}, the stronger correlation between w_i and w_k.

3.2. Trigger Feature Selection

All the tags triggering tag w_i form a set of trigger features of tag w_i in the maximum entropy approach. An example of typical trigger pair (with strong correlation) in music auto-tagging is "falsetto → high pitched". The volume of pairs is too large even if we restrict the trigger pair to two single tags. Furthermore the trigger pairs with low correlation have limited contributions to the contextual modeling actually. In contrast they dilute other strong tag correlations by introducing noise. Therefore trigger features are selected by removing trigger pairs whose correlation values are below the threshold θ.

The selected trigger pair $(w_k \to w_i)$ is reformulated as a constraint in maximum entropy approach in equation (2)

$$f(w_k, w_i) = \begin{cases} 1, & if\ C_{ik} > \theta \\ 0, & otherwise \end{cases} \tag{2}$$

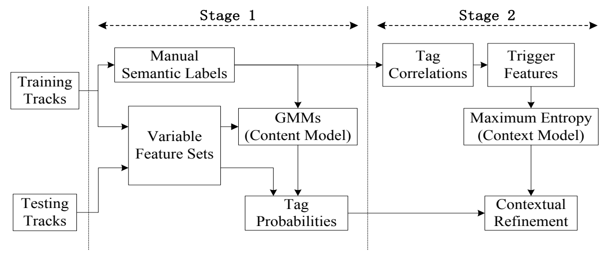

Figure 1. A schematic diagram of the proposed two-stage music auto-tagging system.

Category	Triggered Tag (y)	Trigger Tags(Features) ($X_y = [w_1, w_2, \dots, w_k, \dots]$)
Genre	country	bluegrass, country(best), acoustic guitar, not arousing, folk
	hard rock	angry, aggressive, metal hard rock (best), screaming, not pleasant
Usage	at a party	very danceable, exciting, arousing, catchy, fast tempo
	going to sleep	calming, tender, laid back, not heavy beat, not fast tempo
Solo	Piano	bebop, jazz, violin, cool jazz, touching
	acoustic guitar	falsetto, sleeping, bluegrass, soft rock(best), not recorded

Table 1. TOP-5 trigger features of some "weakly labelled tags"

where C_{ik} represents the correlation of trigger pair $(w_k \to w_i)$ and θ is a threshold and set empirically to eliminate the low correlation tags . For tag w_i, the corresponding trigger features C_{ik} are ranked in a descending order. Top-K features form the trigger feature vector: $X_{w_i} = [w_1, w_2, \dots, w_k, \dots], w_k \in W, k = 1,2, \dots, K$ once θ is set. Table 1 shows some examples of trigger features when $K = 5$ for several "weakly labelled tags" in CAL500.

4. CONTEXT MODELING

The main advantage of maximum entropy (ME) approach lies in its flexibility, which allows stochastic rule systems to be augmented with rich representations [14]. In this section the context models with ME principle are established. The trained context model is applied to refine the probabilities from the first stage of auto-tagging.

4.1. Construction of context models

Assuming that each context model is a conditional probability distribution $p(y|x)$, $y \in W$ is the triggered tag whose probability needs to be refined and $x \in X_y \subseteq R^n$ is the trigger feature (tag) obtained in Section 3, where $X_y = [w_1, w_2, \dots, w_k, \dots]$, $w_k \in W$. The goal of context modeling is to set up a robust model $p(y|x)$ that best accounts for the training data, using the partial evidence from the trigger features. For triggered tag y and the corresponding trigger features in X_y, the training set is $\mathcal{D} = \mathcal{D}_y \cap \mathcal{D}_{\bar{y}}$,where $\mathcal{D}_y = \{(x_1, y), (x_2, y), \dots, (x_i, y), \dots\}$ includes the total tracks labelled by the triggered tag y and $\mathcal{D}_{\bar{y}}$ is the rest of tracks which are not labelled by tag y.

The way to combine trigger features is to "weight" the features in a log-linear or exponential model as a set of constrains [9]:

$$p(y|x) = \frac{1}{Z(x)} exp[\sum_{k=1}^{K} \lambda_k f_k(x,y)] \quad (3)$$

$$Z(x) = \sum_y exp[\sum_{k=1}^{K} \lambda_k f_k(x,y)] \quad (4)$$

where $p(y|x)$ is the conditional probability of predicting an triggered tag y given the seen trigger feature x, and $f_k(x,y)$ is the k^{th} constraint function that maps the trigger pair $(x \rightarrow y)$ to the binary contextual predication, defined by equation (2). λ_k is the weight. K is the total number of constraint functions and $Z(x)$ is a normalization factor that ensures $\sum_y p(y|x) = 1$.

In the conditional maximum entropy framework, the optimal solution $p^*(y|x)$ is the most uncertain distribution that satisfies K constrains over feature expectations:

$$p^*(y|x) = \max_{p \in \mathcal{C}_1} H(p) \quad (5)$$

$$H(p) = -\sum_{x,y} \tilde{p}(x)\, p(y|x) log p(y|x) \quad (6)$$

$$\mathcal{C}_1 \equiv \{p | E_p(f_k) = E_{\tilde{p}}(f_k), k = 1,2,\dots,K\} \quad (7)$$

$$E_{\tilde{p}}(f_k) = \sum_{x,y} \tilde{p}(x,y) f_k(x,y) \quad (8)$$

$$E_p(f_k) = \sum_{x,y} \tilde{p}(x) p(y|x) f_k(x,y) \quad (9)$$

where $H(p)$ denotes the conditional entropy and $E_p(f_k)$ is the expectation of f_k while $E_{\tilde{p}}(f_k)$ is the observed expectation of f_k. $\tilde{p}(x,y) = (v(x,y))/N_{\mathcal{D}}$ is the empirical joint probability of trigger pairs and $\tilde{p}(x) = (v(x))/N_{\mathcal{D}}$ is the empirical marginal probability distribution of the input triggers and $N_{\mathcal{D}}$ is the total number of trigger pairs in the training set $\mathcal{D}$. $v(x,y)$ and $v(x)$ are the occurrence number of $(x \rightarrow y)$ and x respectively.

4.2. Parameter Estimation

Given a ME model with K constraints, the weights $\lambda = [\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_k, \dots]$ for all trigger features can be obtained by maximum likelihood estimation for a best fit given the dataset. The parameter estimation problem then could be translated into an optimization problem as follows.

$$p^*(y|x) = \min_{p \in \mathcal{C}_2} L(p) \quad (10)$$

$$\begin{aligned} L(p) &= \sum_{x,y} \tilde{p}(x,y)\, log p(y|x) \\ &= -H(p) + \sum_{k=1}^{K} \lambda_k (E_{\tilde{p}}(f_k) - E_p(f_k)) \end{aligned} \quad (11)$$

$$\mathcal{C}_2 \equiv \{p | p(y|x) = \frac{exp[\sum_{k=1}^{K} \lambda_k f_k(x,y)]}{\sum_y exp[\sum_{k=1}^{K} \lambda_k f_k(x,y)]}\} \quad (12)$$

where $L(p)$ is the conditional log likelihood of the training set $\mathcal{D}$ and $p^*(y|x)$ is the optimal probability distribution according to the maximum likelihood criterion. It is equivalent to maximum entropy parameter estimation over the set of consistent models [9] as indicated in equation (13)

$$p^*(y|x) = \min_{p \in \mathcal{C}_2} L(p) = \max_{p \in \mathcal{C}_1} H(p) \quad (13)$$

It is suggested in [14] that Limited-Memory BFGS (L-BFGS) is the most effective parameter estimation method for iterative refinement of maximum entropy (ME) models, much better than Generalized Iterative Scaling (GIS) and Improved Iterative Scaling (IIS). Hence L-BFGS algorithm is employed in this work.

4.3. Probability Refinement by the Context models

To keep consistent with the baseline system in [7], $P_q = \{p_q(i)\}$, $i = 1,2,\dots,|W|$ is the quantized probability of the first stage content based auto-tagging, ranging between 0 and 1. The trigger based context models are used to refine the output probabilities P_q of those "weakly labelled tags", which are identified via TF-IDF scores $S_{TF-IDF}(w_i, s_j)$. TF-IDF scores are calculated as a combination of the frequency df_{w_i} of tag w_i and its strength $tf_{w_i s_j}$ over track s_j as shown in equation (14).

$$S_{TF-IDF}(w_i, s_j) = (1 + log(tf_{w_i s_j}))(log \frac{n}{df_{w_i}}) \quad (14)$$

$S_{TF-IDF}(w_i)$ is then denoted as the average of $S_{TF-IDF}(w_i, s_j)$ over all tracks annotated by tag w_i. It is argued that tag w_i has a low df_{w_i} but a high tag probability $tf_{w_i s_j}$ if few tracks are annotated by tag w_i. So the score $S_{TF-IDF}(w_i, s_j)$ is relatively high. Those tags are often sparsely distributed and crucial to the context modeling. A threshold β is set for all tags. If $S_{TF-IDF}(w_i) \geq \beta$, w_i is identified as a weakly labelled tag. Fewer weakly labelled tags are found in "Emotion" and "Song" than other categories.

5. EXPERIMENTS AND RESULTS

The data set employed in this work is the CAL500, which includes 500 tracks and each is represented by 174 "musically relevant" semantic tags spanning six semantic categories [2]. The data set of 450 tracks for training content model is used for trigger-based context model buildup in a fashion of 10 fold cross validation, in which 50 of 450 tracks are randomly selected test sets for context modeling. Two set of experiments are conducted to evaluate the efficiency of context models. One is to evaluate the trigger accuracy of the context models and the other is to compare its performance with the previous system based on the content models only [7] and referred models in [2].

For a triggered tag y, the trigger features given a track is $X_y(s_j) = [w_1(s_j), w_2(s_j), \dots, w_k(s_j), \dots]$. s_j is the j^{th} test track and $w_k(s_j)$ is the k^{th} trigger feature given track s_j. The trigger probability $p(y|w_k(s_j))$ for each trigger pair $(w_k(s_j) \rightarrow y)$ can be obtained from the optimal probability distribution $p^*(y|x)$, $x \in X_y$.The context probability of tag y is defined as:

$$p(y|X_y(s_j)) = \prod_{k=1}^{K} p(y|w_k(s_j)) \quad (15)$$

The trigger accuracy rate of each context model is defined as:

$$Rate_{ME} = N_{co-tri}/N_{test} \quad (16)$$

where N_{co-tri} is the number of tracks in which tag y is "correctly" triggered by the trigger features. N_{test} is the total number of tracks in the test set. If the context probability $p(y|X_y(s_j)) \geq 0.5$ and the softAnnotation of y given track $s_j \geq 0.5$, the trigger is successful.

Category	K=30	K=50	K=80	K=110	K=173 (all)
Emotion	**99.73%**	99.04%	94.95%	92.80%	92.21%
Genre	59.42%	74.29%	91.39%	**97.87%**	93.52%
Instrument	50.58%	65.83%	90.63%	**97.42%**	86.92%
Song	92.01%	**92.57%**	91.86%	91.20%	86.77%
Usage	49.80%	66.27%	87.20%	91.53%	**92.20%**
Vocals	45.33%	60.69%	89.06%	**97.13%**	90.31%

Table 2. The Comparison of Trigger Accuracy Rates of Top-K Trigger Features in a Variety of Tag Categories.

Table 2 shows the average precision rate of the proposed context models for each tag category with different sizes of trigger features invovled. The accuracy of the contextual model is around 90% for all the categories, implying that applying ME approach to obtain the pattern of contextual relationships is in a sense of feasibility. It is also note that increasing of the size of trigger feature vector does not always means a lift in the average accuracy rate, e.g. in 'song' category. Therefore it is preferable to find the suitable K according to the selected threshold θ for each category before establishing the context models as mentioned in Section 3.2. Tags in "Emotion" and "Song" categories acquire a smaller size of trigger features but give higher accuracy comparing to other four categories. It could be explained by the fact that tags in "Emotion" and "Song" are more correlated while tags in the other categories are more independent [13]. Overall the average precision rate of the context model could be enhanced by trigger feature selection by 5.83% or so .

2nd experiment is to evaluate the performance of the proposed auto-tagging system with and without trigger-based probability refinement in terms of three standard information retrieval metrics: average per tag precision ($Mean\_Precision$), average per tag recall ($Mean\_Recall$) and F-score:

$$precision(w_i) = N_{co-anno}/N_{model} \quad (17)$$

$$recall(w_i) = N_{co-anno}/N_{anno} \quad (18)$$

$$Mean\_Precision = \sum_{i=1}^{|W|} precision(w_i) \quad (19)$$

$$Mean\_Recall = \sum_{i=1}^{|W|} recall(w_i) \quad (20)$$

$$F-score = \frac{2\times Mean\_Precision \times Mean\_Recall}{Mean\_Precision + Mean\_Recall} \quad (21)$$

where $N_{co-anno}$ is the number of tracks for which tag w_i is "correctly" annotated by the models (namely GMM(VFS) and GMM+ME, before and after adding trigger-based context refinement respectively). N_{model} is the number of all tracks annotated with tag w_i by the model and N_{anno} is the number of tracks being annotated with tag w_i in soft annotations. $W = \{w_i\}$, $i = 1,2,\ldots,|W|$ is the tag sets in each category. Each test track is annotated with a fixed number of tags, which is defined as "annotation length" A. For comparison purpose, the annotation length A is set same as [2] for each category and the results are listed in Table 3.

It is clear from Table 3 that the system with both and content and context models (GMM+ME) gives better performance than the content model only system (GMM(VFS)) in [7] and the baseline MixHier system in [2].

Category	$A/\|W_c\|$	Model	Mean_ Precision	Mean_ Recall	F-score
Emotion	4/36	MixHier	0.424	0.195	0.267
		GMM(VFS)	0.391	0.152	0.219
		GMM+ME	0.395	0.153	0.220
Genre	2/31	MixHier	0.171	0.242	0.200
		GMM(VFS)	0.207	0.200	0.202
		GMM+ME	**0.229**	0.224	**0.226**
Instrument	4/24	MixHier	0.267	0.320	0.291
		GMM(VFS)	0.227	0.233	0.229
		GMM+ME	0.228	0.233	0.230
Solo	1/9	MixHier	0.060	0.261	0.098
		GMM(VFS)	0.060	0.201	0.093
		GMM+ME	**0.066**	0.214	**0.099**
Usage	2/15	MixHier	0.122	0.264	0.167
		GMM(VFS)	0.227	0.229	0.128
		GMM+ME	**0.287**	**0.267**	**0.276**
Vocals	2/16	MixHier	0.134	0.335	0.191
		GMM(VFS)	0.185	0.224	**0.202**
		GMM+ME	**0.189**	0.221	0.167
All	10/174	MixHier	0.265	0.158	0.198
		GMM(VFS)	0.270	0.173	0.211
		GMM+ME	**0.272**	**0.174**	**0.212**

Table 3. A Comparison of Performance of Music Autotagging system. A is the annotation length and $|W_c|$ is total number of tags in that category. MixHier is the benchmark model proposed by Turnbull [2]. GMM(VFS) is based on variable feature sets and GMM models [7]. GMM+ME is the proposed two-stage framework with GMM(VFS) as the content model and ME as the context model.

The system with GMM+ME model outperforms MixHier in terms of average precision rate and F-score but fail in average recall rate. Though a good balance of precision and recall is always desirable, it has been argued that precision is more important for retrieval and recommender system to maintain the cohesive users. "Weakly labeled tags" in "Genre", "Solo" and "Usage" benefits a lot from trigger-based context modeling with 10.63%, 10% and 26.43% rise in the average per-tag precision and 12%, 6.47% and 16.59% rise in the average per-tag recall after applying context models respectively. It comes to a conclusion that trigger based context modeling by ME approach could improve performance for tags that are poorly represented by content models.

6. CONCLUSION

This paper proposes a music auto-tagging system in a two-stage framework including the content model in our previous work [7] and trigger based context models. The former equipped with the variable feature subsets outputs the probabilistic semantic annotation of tracks and the latter refines the output probabilities of content models by modeling contextual information between tags via ME approach with selected trigger features. The system with combined content and context models improves by 0.74% and 0.58% over the content model only system while enjoys 2.64% and 10.12% higher than the baseline in terms of precision and recall rates. Weakly labeled tags in "Genre", "Solo" and "Usage" which are annotated by the content models have been improved around 10% in precision rates.

7. REFERENCES

[1] P. Knees, M. Schedl, “Music similarity and retrieval,” in proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, 1125-1125

[2] D. Turnbull, L. Barrington, D. Torres, et al, “Semantic annotation and retrieval of music and sound effects,” IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 467-476, 2008.

[3] F. Charte, A. Rivera, M.J.D Jesus, et al. “A first approach to deal with imbalance in multi-label datasets,” Springer Berlin Heidelberg, 150-160, 2013.

[4] S. Ness, A. Theocharis, G. Tzanetakis, et al, “Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs,” in proc. ACM, 2009, 705-708.

[5] Y. Yang, Y. Lin, A. Lee, et al, “Improving musical concept detection by ordinal regression and context fusion,” in proc. ISMIR, 2009, 147–152.

[6] R. Miotto, G. Lanckriet, “A generative context model for semantic music annotation and retrieval,” IEEE Transactions on Audio, Speech, and Language Processing, 20(4): 1096-1108, 2012.

[7] J. J. Yin, Q. Yan, Y. Lv, et al, “Music Auto-tagging with Variable Feature Sets and Probabilistic Annotation,” in proc. CSNDSP, 2014, 156-160.

[8] A. L. Berger, V. J. D. Pietra, S. A. D. Pietra, “A maximum entropy approach to natural language processing,” Computational linguistics, 22(1): 39-71, 1996.

[9] A. Ratnaparkhi, “Maximum entropy models for natural language ambiguity resolution,” Ph.D. thesis, University of Pennsylvania, 1998.

[10] C. Argyropoulos, A. Daskalakis, G. C. Nikiforidis, et al, “Background adjustment of cDNA microarray images by Maximum Entropy distributions,” Journal of Biomedical Informatics, 43(4): 496-509, 2010.

[11] G. Tsatsaronis, N. Macari, S. Torge, Dietze H, et al, “A Maximum-Entropy approach for accurate document annotation in the biomedical domain,” Journal of Biomed Semantics, 3(Suppl 1): S2, 2012

[12] R. Lau, R. Rosenfeld, and S. Roukos, “Trigger-based language models: a maximum entropy approach,” in proc. ICASSP, 1993, 2: 45–58.

[13] B. Markines, C. Cattuto, F. Menczer, et al, “Evaluating similarity measures for emergent semantics of social tagging,” in proceedings of the 18th international conference on World wide web, ACM, 2009, 641-650.

[14] R. Malouf, A. Informatica, R. Groningen, “A comparison of algorithms for maximum entropy parameter estimation,” in proceedings of the 6th conference on Natural language learning, 2002, 20: pp.1-7.