Deriving Music Theme Annotations from User Tags

Kerstin Bischoff, Claudiu S. Firan, Raluca Paiu*
L3S Research Center / Leibniz Universität Hannover
Appelstr. 9a
30167 Hannover, Germany
{bischoff,firan,paiu}@L3S.de

ABSTRACT

Music theme annotations would be really beneficial for supporting retrieval, but are often neglected by users while annotating. Thus, in order to support users in tagging and to fill the gaps in the tag space, in this paper we develop algorithms for recommending theme annotations. Our methods exploit already existing user tags, the lyrics of music tracks, as well as combinations of both. We compare the results for our recommended theme annotations against genre and style recommendations – a much easier and already studied task. We evaluate the quality of our recommended tags against an expert ground truth data set. Our results are promising and provide interesting insights into possible extensions for music tagging systems to support music search.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Collaborative Tagging, High-Level Music Descriptors, Theme Tag Recommendations, Metadata Enrichment

1. INTRODUCTION

Currently no available search engine supports music search by sample music files, thus people are still constrained to search for music using textual queries. In this context, supporting users in providing meaningful tags for music tracks becomes crucial – tags and other metadata (*e.g.* extracted from ID3 tags) can be indexed and later be used to support music search. [1] discusses the problems arising from the gaps existing between the types of tags employed by users for tagging music data and the types of keywords used for music search: theme-related words represent 5% of the tags for songs, though when searching for music, 30% of the queries are theme-related (*e.g.* "halloween party music"). One possibility to make users use keywords from the categories we need is to unobtrusively recommend such tags and thus support them in the tagging process. Specifically, our goal is to support users by recommending tags referring to usage context information ("themes"). The "theme of a song" refers to the context or situation which fits best when listening to the music track, *e.g. at the beach*, *dinner ambiance*, *party time*, *road trip*, *etc.*Within an application recommendations could be presented to the user, who can select the relevant ones and add them to the music track. Recommended theme tags can also be indexed to enrich the metadata for a specific track and to automatically create theme-based playlists.

2. RELATED WORK

Several existing papers aim at automatically inferring additional information from available content or (user generated) metadata. There are approaches that exploit user generated tags to recommend additional tags for pictures [6], for blog posts [8] or personalized tags for Web pages [2]. For music, enrichment recently focuses on deriving mood information based on extracted accoustic data [3, 5, 7]. Establishing a basis for this, [4] studies the relationships between moods and artists, genres and usage metadata. In contrast to existing music metadata enrichment studies, we rely only on social tags and no low-level features of the tracks.

3. DATA SET DESCRIPTION

For our experiments, we collected the *AllMusic.com* pages covering music themes, genres and styles. We found 73 themes, 20 genres and 633 styles (more fine-grained music genre classes). From the pages corresponding to themes / genres / styles, we also gathered information related to which music tracks fall into these categories. With this procedure, we ended up with 13,948 songs. Looking at the songs identified in each of the categories, we have 1,164 track-theme, 1,521 track-genre, and 16,023 track-style assignments. For all these songs we also crawled the tags associated to these music tracks by users in *Last.fm*. We have 81,964 different tags with their usage frequency. To investigate whether lyrics can provide added value in the task of theme and genre / style recommendation, we also obtained the lyrics for the 13,948 tracks from *www.lyricsdownload.com* and *www.lyricsmode.com*, if available.

4. RECOMMENDATION ALGORITHMS

For recommending themes, we base our solution on collaboratively created social knowledge – *i.e.* tags associated to music tracks – extracted from *Last.fm*, as well as on lyrics

*work supported by the PHAROS project funded by the European Commission (FP6 - IST Contract No. 045035)

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

information. Based on already provided user tags, on the lyrics of the music tracks, or on combinations of the two, we built classifiers which try to infer other annotations corresponding to themes appropriate for the songs. For comparison reasons, we additionally experiment with predictions for music genres and styles. For building these classifiers, we use the open source machine learning library Weka[1]. In the experiments presented in this paper, we use the Naïve Bayes Multinomial implementation available in Weka. We have one classifier trained for the whole available set of classes (*i.e.* either for themes or genres or styles). This classifier produces for every song in the test set a probability distribution over all classes (*e.g.* over all themes). Thus, one or more classes (based on probabilities or on a given rank number) can be then assigned to each song.

We experiment with three types of input features for the different sets of classifiers: (1) tags; (2) words from lyrics; or (3) tags *and* words from lyrics. Depending on the type of features used to train the classifier and on the type of class that the classifier will assign to songs, we propose 9 methods (3 types of output classes – themes, genres, styles – and 3 types of features – tags, lyrics, tags+lyrics). Each of these algorithms uses a different number of input features, as the sets of *AllMusic* songs having theme, genre or style labels do not overlap perfectly. We experimented with feature selection (*e.g.* Information Gain) but results showed that the full set, though introducing noise, is better suitable for learning.

As we need a certain amount of input data in order to be able to consistently train the classifiers, we discard those classes having less than 30 songs assigned. For each of the three types of classes, a classifier learns a model based on the presented features. Then the model is applied to any new, unseen data. We can choose how many tags are recommended to the user based on the probabilities resulted from the classification or by setting an absolute threshold.

5. EVALUATION & CONCLUSIONS

The evaluation we perform aims to automatically measure the quality of our tag prediction algorithms. As ground truth data we use the *AllMusic.com* data set, on which we perform 10-fold cross-validation. Given a *Last.fm* music track, we predict possible theme / genre / style annotations and compare our output against the manual assignments of *AllMusic* experts for the same song. We present the results for all our experimental runs in Table 1. The evaluation metrics we analyze are Hit-Rate at rank k (H@3, H@5), showing whether a good descriptive tag is contained among the top-k recommended tags; R-Precision (RP), precision at the total number of relevant tags; and Mean Reciprocal Rank (MRR). We concentrate on the H@3 metric, as we recommend three annotations to the users to choose from. We consider three annotations a good compromise, providing enough suggestions and at the same time not overwhelming the users with too much information.

We observe that the best performing methods are those using tags as input features for the classifiers. When combining tags and lyrics as features, the corresponding methods perform much better than those based only on lyrics and they sometimes also slightly outperform the tag-based methods. Lyrics, in contrast to tags, introduce noise, as many song texts contain all sorts of interjections (*e.g.* “hey”, “oh”, “uh-huh”, *etc.*), slang or simply informal English. As expected, the best results we obtain are for the genre-tag recommendations: H@3 of 0.97 for the case of tags as features. Styles do not perform as good as genres (H@3 of 0.76), mostly due to the fact that the *AllMusic* labels are too fine-grained to clearly distinguish between them. Given the difficulty of agreeing on a single, appropriate music genre taxonomy, some of these fine distinctions may also be worth discussing. For the case of theme recommendations, the best results, H@3 of 0.80, are achieved for the algorithm both using only tags or a combination of tags and lyrics as features.

Table 1: Overall experimental results

	Features	H@3	H@5	RP	MRR
Theme	Tags	0.80	0.92	0.49	0.67
	Lyrics	0.56	0.72	0.26	0.46
	Tags+Lyrics	0.80	0.94	0.48	0.67
Genre	Tags	0.97	0.98	0.83	0.91
	Lyrics	0.85	0.93	0.60	0.75
	Tags+Lyrics	0.93	0.98	0.76	0.86
Style	Tags	0.76	0.85	0.48	0.65
	Lyrics	0.22	0.29	0.10	0.21
	Tags+Lyrics	0.62	0.72	0.37	0.54

The results indicate a good performance of our algorithms in correctly recommending themes or genre / style annotations. Given the self-reinforcing nature of user generated tags, the set of recommended theme tags will not only enrich our future training set for learning, but will probably also enable fully automatic theme tag assignment without user interaction. Using our approach, music becomes searchable by associated themes, providing a first step towards effectively searching music by textual descriptive queries.

For the future we plan to investigate this issue further, as well as mood classification using metadata and how to improve feature selection by automatic identification of tag types (*e.g.* Topic, Author, Usage context). Merging our approach with content-based methods trying to solve the same task is also worth examining.

6. REFERENCES

[1] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *CIKM*, 2008.

[2] A. Byde, H. Wan, and S. Cayzer. Personalized tag recommendations via tagging and content-based similarity metrics. In *ICWSM*, 2007.

[3] Y. Feng, Y. Zhuang, and Y. Pan. Popular music retrieval by detecting mood. In *SIGIR*, 2003.

[4] X. Hu and J. S. Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *ISMIR*, 2007.

[5] D. Liu, L. Lu, and H.-J. Zhang. Automatic mood detection from acoustic music data. In *ISMIR*, 2003.

[6] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, 2008.

[7] J. Skowronek, M. McKinney, and S. van de Par. A demonstrator for automatic music mood estimation. In *ISMIR*, 2007.

[8] S. Sood, S. Owsley, K. Hammond, and L. Birnbaum. Tagassist: Automatic tag suggestion for blog posts. In *ICWSM*, 2007.

[1] `http://www.cs.waikato.ac.nz/~ml/weka`