From Lyrics to Their Interpretations: Automated Reading between the Lines

Kahyun Choi
GSLIS
University of Illinois
MC-493, Suite 329
Champaign, IL 61820
+1.217.333.3282
ckahyu2@illinois.edu

ABSTRACT

Although musical metadata, such as genre, mood, and usage, have been widely studied, there is a less explored type of metadata, the subject of the lyric, which might be potentially useful in music digital libraries. In my previous research, I have explored the use of the subject metadata in the context of music classification based on it. The biggest challenge of the classification task is that lyrics are often hard to analyze automatically due to their poetic nature. What I discovered is that user-created interpretations such as those readily available on songmeanings.com are more suitable for this kind of task than the lyric text itself. A subsequent project was to apply Latent Dirichlet Allocation (LDA) topic modeling on this collection of interpretations, and then to automatically discover the underlying subjects. This unsupervised topic modeling technique is especially meaningful, because human annotation of the subject categories per each song can be difficult and expensive, while the interpretation can be seen as a verbal explanation of the subject categories. I proposed a topic refinement system that uses some intrinsic topic evaluation techniques. The system is based on Normalized Pointwise Mutual Information (NPMI) and topic weights that are available as the results of LDA, followed by an extrinsic evaluation method as well to demonstrate the quality of refinement. The next phase of this study is to learn the relationship between the lyric and the interpretation. Although I found the interpretation dataset to be especially useful for the subject discovery and prediction, it is true that the interpretation of lyrics has certain limitations such as its reliance on volunteer work and its scarcity. I believe that an elaborately designed intelligent system should be able to predict the subject from nothing but the lyric texts. However, the relationship between the lyrics and existing interpretations requires a development of an intelligent mapping technique. I plan to learn the mapping by using LDA as a dimensionality reduction technique, and then learning a regression model between the dimension-reduced lyrics and interpretations, which is a universal way to learn all the relationships that I need.

JCDL'15, June 21 - 25, 2015, Knoxville, TN, USA.
Copyright 2010 ACM 1-58113-000-0/00/0010 …$15.00.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: indexing methods

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Topic Models, Music Digital Library, Interpretations of Lyrics, Text classification

1. INTRODUCTION

In the Music Digital Libraries (MDL), subject metadata is very important in searching and browsing collections. Because it is expensive to hire human evaluators to annotate a big music database, the need for an automatic subject metadata enrichment method is increasing.

There are a few terms used to try to define this concept, such as “subject,” “topic,” “theme,” and “aboutness.” They all interchangeably refer to what the lyric is about, but the usage varies across studies. For example, “themes” are used to refer to the events where the music is used, or the mood of the music, e.g., “party music,” “wedding songs,” “mellow music” [1]. “Topic” is another term for the similar purpose, but it mostly applies in cases where topic modeling algorithms are used. In [2] “theme” is indeed used to explain what a song is about, in place of the term “subject” I use here in this work. Meanwhile, Fairthorne [3] defines an intentional “aboutness” of a text as its topic, which is a definition that goes well with mine. While recognizing those slightly different definitions, in this work I use the term “subject” and “topics” interchangeably to refer to the concept.

As for the inputs to the methods, first of all, we can think of the lyric itself to determine what this song is about. However, it is often difficult for even humans to understand it because of the poetic nature of lyrics. For this reason, the lyric-only input has inherent limitations for the task. Instead, in this work I focus on an auxiliary dataset, where people’s interpretations for a particular lyric are available in the form of comments. For example, there is a website called *songmeanings.com* [1], where different understandings of songs are shared and rated by the users. There are already more than 1,700,000 interpretations in the service and the number is growing. I started off with the knowledge that interpretation of lyrics would be more suitable than the lyrics alone for the automatic subject analysis because interpretation is straightforward while lyrics are often ambiguous. Therefore, there

[1] http://songmeanings.com/

can be three different combinations of the inputs, i.e. lyrics, interpretations, and the combination of the two.

One way to automatically assign subject labels is to build a topic classification system that takes the text sources as input samples. Widely used inputs are metadata, such as genre, mood, and composers to songs, whereas we focus on lyrics and their interpretations. Once we have collected those input samples with corresponding topic categories as their ground-truth output pairs, the classification system can be trained to learn the relationship between text sources and topic labels, and then to predict a right label for an unknown song. On top of its original use as a predictor, this supervised system can also judge the usability of the different text input sources based on the classification performances.

While classification assumes that a song is associated with one particular subject category, we can use topic modeling to assign multiple subjects to each song in an unsupervised manner. Classification is suitable when it comes to a fixed set of output labels, but most of the time it is not well defined how many subject categories there are in a big music library. In the latter case, topic modeling is more appropriate particularly to explore and discover unknown number of subjects. This unsupervised method has been widely used in detecting topics from various text collections ranging from academic papers to tweets. However, since learned topics are mixtures of useful topics and junk topics, it is important to devise some techniques to sort out the topics in the order of their quality for the further use in the music information retrieval systems as a browsing and searching option.

In my prior work [4], I showed that the interpretations available in *songmeanings.com* are more useful than lyrics themselves when it comes to classifying lyrics based on their underlying subjects. Also, I have worked on an unsupervised system to discover subject labels from the interpretations automatically [5]. On this system I used Latent Dirichlet Allocation (LDA) to learn some candidate subjects from the interpretations, and came up with a method to filter out noisy ones in some systematical ways based on topic weights and intrinsic topic coherence.

Even if interpretations are better sources than lyrics, there is a scalability issue as not every song has a corresponding set of user interpretations. Therefore it is necessary to find a way to extract the subject information only from lyrics as well, for instance by utilizing topic modeling results on the training set. If there is strong relationship between lyrics and interpretations and a way to model the relationship, it might be also possible to generate interpretations of currently un-interpreted lyrics using the model.

To sum up, the following research questions will be answered in the dissertation.

RQ1: Can we exploit lyrics and interpretations in automatically capturing subject information of music? Which one is more useful in the task?

RQ2: How can we use topic modeling user-generated interpretations to assign subject metadata to music?

RQ3: How can we assign subject metadata to lyrics without any interpretation by exploiting the relationship between lyrics and their interpretations?

2. RELATED WORK

When music is accompanied by lyrics, the subject, or what the lyric is about, is of great interest to music listeners. User studies support this argument that users rely heavily on the subject metadata: an online survey says that 33.4% of 427 music listeners responded that they would like to use subject to search and browse music [2]. What is noticeable is that the percentage is higher than the other types of metadata, e.g. mood, time period, instruments, etc. It is also interesting to note that around 15% of AOL music queries among sampled 300 queries are about subjects [6]. Bainbridge et al. [7] also showed that among 502 music-related articles in Google Answers, now an obsolete Q&A service, “Lyric Story (storyline of song)” was described as one of the information needs. Therefore, it seems promising to have a Music Digital Library (MDL) that is annotated with subject metadata as well.

Lyrics are the starting point when it comes to music topic identification. They have received a lot of attention from researchers for many reasons [8][9][10]. For instance, most popular songs have lyrics and these are easily collected from the web [8]. In addition, they have been used as features to determine artist similarity [8], music mood [9], music genre [10], etc. by complementing other metadata or audio features. Despite their usefulness in various Music Information Retrieval (MIR) tasks, they have much in common with poetries. It is evidenced by [11] that lyrics and poetries are confused in automatic classification systems. To be specific, more than 40% of lyrics written by some poetic lyricists such as Bob Dylan and Ed Sheeran are misclassified as poetry [11]. To overcome the difficulty of automatic analysis of poetic lyric itself, my work uses user-generated interpretations as well, as they tend to contain clearer explanation of the lyrics.

There have been efforts to extract the subject of the lyric automatically. Mahedero et al. solved this problem by using a naïve Bayes classifier on a small dataset (125 songs) and five subject labels [12]. Bischoff et al. [13] proposed a music annotation system that suggests opinion, usage, genre, and style information based on social tags and lyrics. In [13], the classifier using a combination of tags and lyrics outperformed those using only tags or lyrics when classifying theme and mood. Notably, using lyrics in addition to tags reduced the performance by adding noise in genre and style annotation tasks. This indicates that the usefulness of text sources depends on the type of metadata predicted from them. Hence in my study, I also investigate the usefulness of various text sources and their combinations for predicting the subject information.

Kleedorfer et al. proposed an unsupervised model using Non-negative Matrix Factorization (NMF) as a topic model, which showcases a good match with human evaluations [14]. Another unsupervised topic modeling work, by Sasaki et al., was done by using LDA, which is a more advanced topic modeling technique with an additional Bayesian treatment of the topics with Dirichlet priors on topics [15].

3. DATASET

We mainly use three different services to collect our dataset: *songfact.com*, *songmeanings.com*, and the Million Song Dataset (MSD)[2].

Songfact.com is where we collect ground-truth subject categories, that are determined by “interviews, books, magazines, newspaper

[2] We focus on the common songs in both collections, because we need the metadata in MSD in future work.

articles, reference materials, and publicity releases."[3] Among the 126 subject categories, we selected the most popular ones (6 or 10 depending on the experiment).

Interpretations are from *songmeanings.com*. First, the songs with more than 5 attached comments are selected for more reliable results. Second, among those songs, I only select the ones that also appear in the Million Song Dataset (MSD), a free music collection with useful audio features and metadata. The selected songs amount to 24,436 eventually.

4. METHODOLOGY

4.1 Classification

The goal of classification is to train an autotagging system that is designed to label a song based on the predicted subject. A song will be represented by lyrics, the interpretations, and the combination of lyrics and their interpretations, respectively.

I found that both K-Nearest Neighbors (KNN) classifiers and Support Vector Machines (SVM) provide meaningful results thanks to their non-linear nature of decision boundaries. For clarity, however, in this proposal I only discuss the KNN classifiers without any loss of generality. I want to show that the features from the interpretation data can be better than the ordinary lyric-based ones, and coming up with the best-tuned classifier is not my main interest. Moreover, KNN classifiers are sophisticated enough depending on the choice of K, the number of neighborhood samples that it sees. Plus, they usually perform better than naive Bayes classifiers. I also found that cosine distance is useful as the distance metric.

I also perform some statistical tests to justify the significance of the classification results. Friedman's ANOVA test and the Tukey-Kramer "Honestly Significant Difference" (HSD), that have been commonly used in various MIREX[4] tasks [16], are adopted in my research as well. They were also used to find the most useful features in the music mood classification task [9].

4.2 Topic Modeling

Aside from the classification experiments, I also investigate whether there is an optimal number of topics, and how those topics contribute to each lyric example. This is a pretty standard application of topic modeling as a topic discovery tool [17][18][19]. I learn two probability distributions: one describes word distribution per a topic, and the other describes contribution of topics per document. This way, I can handle the case when a lyric contains multiple subjects, as opposed to the classification case or a clustering technique. LDA is a powerful generative model thanks to its Dirichlet prior assumptions over the parameters [17]. I particularly focus on the topic parameters, as *a priori* for the topic distributions per each document to judge the popularity of a given topic. Specifically, the Dirichlet parameters, each of which corresponds to a topic, are to show global popularity of topics in the dataset. In general, the bigger the Dirichlet parameters, the popular the topics are.

I also evaluate the topic modeling results in both intrinsic and extrinsic ways [18]. First, I use Pointwise Mutual Information (PMI), proposed by Newman et al. [18] to better correlate the results with human assessment on topic coherence. The reliable performance of this evaluation technique comes from the fact that it calculates the PMI of pairs of terms in the computed based on their co-occurrence in a large text database, such as Wikipedia. I specifically used a Normalized PMI (NPMI), an implementation available in the Palmetto online tool [19]. NPMI values are between -1 and 1 by the use of the normalization factor, $-\log p(w_i, w_j)$:

$$NPMI(w_i, w_j) = \frac{\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i,w_j)}.$$

Here, -1 is an extreme case where the terms never occur together, 0 means independence, and 1 is for the other extreme where they always occur together.

Extrinsic evaluation is done by a ground-truth test set extracted from *songfacts.com* of 422 songs manually labeled with the six most popular subjects: *heartache*, *sex*, *parents*, *religion*, *drugs*, and *war*. I use a majority-voting approach to coupling LDA results and the *songfacts.com* labels: for each song, the top three LDA topics are mapped to the *songfacts.com* categories.

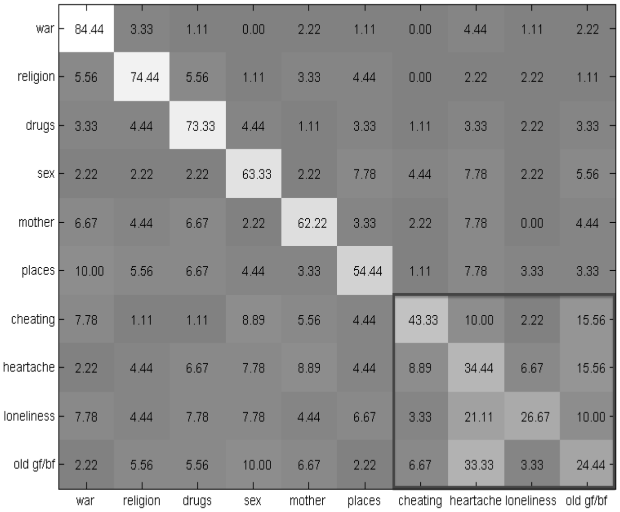
5. PROGRESS AND FUTURE WORK

5.1 Automatic subject classification (RQ1)

In this work I compared the previously unexplored form of user data to another conventional textual source, i.e. interpretations versus lyrics, at subject prediction tasks. First, for the 900 songs used for the comparison of the different source types, it is shown that the case with only the interpretation performed the best (Table 1). Friedman's ANOVA test on the classification accuracies says that the case with interpretations is significantly more accurate than lyrics or the combination of the two ($p < 0.05$).

Table 1. Classification accuracies of different sources and TFIDF weighting method

Source	Accuracy	
	TF	TFIDF
Interpretations	46.11%	**54.11%**
Lyrics	22.78%	26.00%
Interpretations + Lyrics	43.33%	47.44%



	war	religion	drugs	sex	mother	places	cheating	heartache	loneliness	old gf/bf
war	84.44	3.33	1.11	0.00	2.22	1.11	0.00	4.44	1.11	2.22
religion	5.56	74.44	5.56	1.11	3.33	4.44	0.00	2.22	2.22	1.11
drugs	3.33	4.44	73.33	4.44	1.11	3.33	1.11	3.33	2.22	3.33
sex	2.22	2.22	2.22	63.33	2.22	7.78	4.44	7.78	2.22	5.56
mother	6.67	4.44	6.67	2.22	62.22	3.33	2.22	7.78	0.00	4.44
places	10.00	5.56	6.67	4.44	3.33	54.44	1.11	7.78	3.33	3.33
cheating	7.78	1.11	1.11	8.89	5.56	4.44	43.33	10.00	2.22	15.56
heartache	2.22	4.44	6.67	7.78	8.89	4.44	8.89	34.44	6.67	15.56
loneliness	7.78	4.44	7.78	7.78	4.44	6.67	3.33	21.11	26.67	10.00
old gf/bf	2.22	5.56	5.56	10.00	6.67	2.22	6.67	33.33	3.33	24.44

Figure 1. Subject classification confusion matrix using interpretations presented in accuracy rank order

[3] http://www.songfacts.com/about.php

[4] http://www.music-ir.org/mirex/wiki/MIREX_HOME; Music Information Retrieval Evaluation eXchange is a community-based formal evaluation framework for music information retrieval systems.

It is also shown in Figure 1 that the interpretation-only case is promising in the classification task (the column of the matrix represents predicted classes, while the row represents the actual ground truth classes). There are six classes with higher accuracies: *war*, *religion*, *drugs*, *sex*, *parents*, and *places*. There are four categories with low performances: *old girl/boyfriend*, *loneliness*, *heartache*, and *cheating*. I believe that there is some correlation between those categories, such as the negative moods shared in *loneliness* and *heartache*.

5.2 Topic Modeling Interpretations (RQ2)

In this part, I developed an automatic topic discovery system that takes user interpretations as the input and returns some subject categories that underlie the dataset. I also develop a filtering technique to further refine the resulting topics, which are apt to be noisy. For the experiments, 24,436 popular songs that are common in both *songmeanings.com* and MSD are included. Topics are learned using LDA. The coherence of topics is first evaluated by using the NPMI metric, with the top ten most frequent words in each topic and their co-occurrences in Wikipedia. Also, the topics with Dirichlet parameters that are too low indicate that they are too rare to keep. The procedure is depicted in Figures 2 and 3.

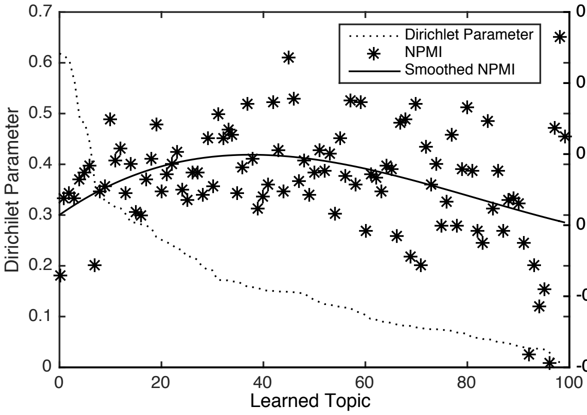


Figure 2. Collection probabilities and NPMI values of LDA topics (k=100).

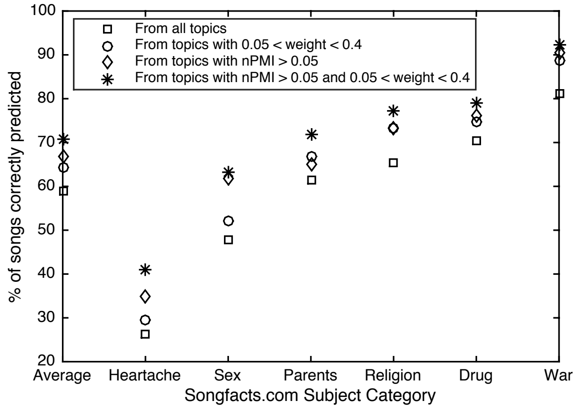


Figure 3. Extrinsic evaluation results (when we assume that top three topics are all equally important).

External evaluation is done on the subset of 422 songs that are manually assigned to six popular *songfacts.com* categories. I found that 71% of the manual assignments are correct, showcasing that the LDA topics on interpretations and the proposed filtering methods are promising for automatically building subject metadata in music digital libraries. I hope that a similar technique can be also used in the collections of poetry and fiction.

5.3 Mapping lyrics and interpretations (RQ3)

Even if the user-generated interpretations are more useful in capturing topic information, most of the songs do not have corresponding interpretations. I believe that this scalability issue can be resolved by inferring topics of a lyric without any interpretation from pre-learned relationship between lyrics and user-generated interpretations.

For this work, the same interpretation dataset of 24,436 popular songs from the previous work will be used in addition to the corresponding lyrics from *songmeanings.com*. First of all, I plan to apply the topic modeling algorithm, LDA, to extract the semantic layers of lyrics and interpretations respectively. Later, mapping functions between dimension-reduced lyrics and interpretations will be devised to take a new lyric as an input and return a possible topic distribution as an output. In order to evaluate the system performance, 10-fold cross validation will be used by splitting lyrics and interpretations into 10 sets of training and test data.

6. ACKNOWLEDGMENTS

I would like to express my appreciation to my doctoral supervisor, J. Stephen Downie for his guidance and support. I am also grateful to Jin Ha Lee and Xiao Hu for giving me invaluable feedback.

7. REFERENCES

[1] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, “How do you feel about dancing queen?: deriving mood & theme annotations from user tags,” *In Proc. of the ACM/IEEE Joint Conf. on Digital Libraries*, 2009, 285-294.

[2] J. H. Lee and J. S. Downie, “Survey Of Music Information Needs, Uses, And Seeking Behaviors: Preliminary Findings,” In *Proc. of 5th Int. Soc. for Music Inform. Retrieval Conf.*, Barcelona, Spain, Oct. 2004, 441–446.

[3] R. A. Fairthorne, “Content analysis, specification and control,” Annual Review of Information Science and Technology 4, 1969, 73-109.

[4] K. Choi, J. H., and J. S. Downie, “What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics,” *In Proc. of the ACM/IEEE Joint Conf. on Digital Libraries*, 2014, 453-454.

[5] K. Choi, J. H., C. Willis, and J. S. Downie, “Topic Modeling Users’ Interpretations of Songs to Inform Subject Access in Music Digital Libraries,” *In Proc. of the ACM/IEEE Joint Conf. on Digital Libraries*, 2015.

[6] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, “Can all tags be used for search?” *In Procc of the ACM Conf. on Information and knowledge management,* 2008, 193-202.

[7] D. Bainbridge, S. J. Cunningham, and J. S. Downie, "How people describe their music information needs: A grounded theory analysis of music queries,” *In Proc. of 4th Int. Soc. for Music Inform. Retrieval Conf.*, 2003, 221-222.

[8] B. Logan, A. Kositsky, and P. Moreno, “Semantic analysis of song lyrics,” *In Proc. of Multimedia and Expo, 2004,* 827-830.

[9] X. Hu and J. S. Downie, “Improving mood classification in music digital libraries by combining lyrics and audio,” *In Proc. of the ACM/IEEE Joint Conf. on Digital Libraries*, 2010, 159-168.

[10] C. McKay, J. A. Burgoyne, J. Hockman, J. BL Smith, G. Vigliensoni, and I. Fujinaga, "Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features," *In Proc. of 11th Int. Soc. for Music Inform. Retrieval Conf.,* Utrecht, Netherlands, Aug. 2010, 213-218.

[11] A. Singhi and D. G. Brown, “Are poetry and lyrics all that different?”, *In Proc. of 15th Int. Soc. for Music Inform. Retrieval Conf.*, Taipei, Taiwan, Oct. 2014, 471-476.

[12] J. P. Mahedero, Á. MartÍnez, and P. Cano, “Natural language processing of lyrics,” In *Proc. of the 13th annual ACM Int. Conf. on Multimedia*, Singapore, Nov. 2005, 475-478.

[13] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, “How do you feel about dancing queen?: deriving mood & theme annotations from user tags,” *In Proc. of the ACM/IEEE Joint Conf. on Digital Libraries,* 2009, 285-294.

[14] F. Kleedorfer, P. Knees, and T. Pohle, “Oh Oh Oh Whoah! Towards Automatic Topic Detection in Song Lyrics,” *In Proc. of 9th Int. Soc. for Music Inform. Retrieval Conf.*, Philadelphia, PA, Sep. 2008, 287-292.

[15] S. Sasaki, K. Yoshii, T. Nakano, M. Goto, and S. Morishima, “LyricRadar: A Lyrics Retrieval System Based on Latent Topics of Lyrics,” *In Proc. of 15th Int. Soc. for Music Inform. Retrieval Conf.*, Taipei, Taiwan, Oct. 2014, 585-590.

[16] J. S. Downie, “The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research,” Acoustical Science and Technology, 29, 4, 2008, 247-255.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *The Journal of Machine Learning Research*, vol. *3*, 993-1022, 2003

[18] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic evaluation of topic coherence,” *In Proc. of Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, LA, CA, June, 2010, 100-108.

[19] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measure,” *In Proc. of the 8th ACM Int. Conf. on Web Search and Data Mining*, 2015.