What is This Song about Anyway?: Automatic Classification of Subject Using User Interpretations and Lyrics

Kahyun Choi
University of Illinois
Champaign, IL 61820
ckahyu2@illinois.edu

Jin Ha Lee
University of Washington
Seattle, WA 98195
jinhalee@uw.edu

J. Stephen Downie
University of Illinois
Champaign, IL 61820
jdownie@illinois.edu

ABSTRACT
Metadata research for music digital libraries has traditionally focused on genre. Despite its potential for improving the ability of users to better search and browse music collections, music subject metadata is an unexplored area. The objective of this study is to expand the scope of music metadata research, in particular, by exploring music subject classification based on user interpretations of music. Furthermore, we compare this previously unexplored form of user data to lyrics at subject prediction tasks. In our experiment, we use datasets consisting of 900 songs annotated with user interpretations. To determine the significance of performance differences between the two sources, we applied Friedman's ANOVA test on the classification accuracies. The results show that user-generated interpretations are significantly more useful than lyrics as classification features ($p < 0.05$). The findings support the possibility of exploiting various existing sources for subject metadata enrichment in music digital libraries.

Categories and Subject Descriptors
H.3.2.d [**Information Storage and Retrieval**]: Metadata.

General Terms
Measurement, Performance, Human Factors

Keywords
Music, Music Information Retrieval, Subject, Metadata, Data mining, Text classification, User-generated content, Lyrics

1. INTRODUCTION
Subject metadata, in addition to descriptive metadata, serve an important role for users browsing and/or searching information in digital libraries. In the Music Information Retrieval (MIR) and Music Digital Libraries (MDL) domains, determining and representing what a song is about has always been a challenge [1],0. The massive amount of digital music and the high cost of human subject annotation call for the development of automatic music subject classification methods. To date, lyrics and tags have been used as sources for feature vectors [3][4]. Beyond tags and lyrics, user-generated information (e.g., music reviews, users' interpretation of songs) has received relatively little attention.

In this poster, we evaluate music subject classification results based on two sources of information: users' interpretations of songs and lyrics. Our objective is to compare the performance of multiple classification algorithms in order to evaluate the usefulness of these sources on classifying the songs based on their subjects. We compare accuracies of classification algorithms with three possible combinations from two types of sources.

2. DATASET
The subject categories and list of songs for each category are collected from songfact.com, a website which provides various song information including what the song is about. According to the website, subject of song is determined based on "interviews, books, magazines, newspaper articles, reference materials, and publicity releases." [1] Among the 126 subject categories, we selected the 10 most popular ones to be used as ground truth for our classification experiment.

Users' song interpretations and lyrics were collected from songmeanings.com[2], a website where music listeners post their understanding and interpretation of what a song is about and discuss with other community members. Only those songs with more than five comments were used in the experiments to get more reliable results. We assume that users' interpretations can often reveal the deeper meaning of the song and/or what the artist intended to convey in addition to what a song is literally about which is generally captured well by lyrics.

The dataset of 900 songs was used to compare lyrics and interpretations regarding their classification performance. Among the 126 songfact subject categories, the top 10 categories (i.e., heartache, places, sex, an old girlfriend or boyfriend, drugs, war, a mother or father, spirituality or religion, loneliness or isolation, and cheating) were selected. In order to have a balanced dataset, we collected the same number of songs with user interpretations for each of these 10 categories. The maximum number of songs with user interpretations we could collect across these categories was 90, thus resulting in the first dataset with 900 songs.

3. EXPERIMENTS

3.1 Data Preprocessing
The first preprocessing step of comments and lyrics was converting ASCII to text and removal of HTML tags. Then, the text stream was tokenized and only the words and digits were saved. Stemming (Porter stemming[3]) was used to address grammatical variations. In order to remove terms with especially

[1] http://www.songfacts.com/about.php
[2] http://songmeanings.com
[3] http://tartarus.org/martin/PorterStemmer/

978-1-4799-5569-5/14/$31.00 ©2014 IEEE

high or low frequency of occurrences, we eliminated stopwords[4] and words that appeared fewer than 5 times. In order to equally compare the two sources, we did not apply additional preprocessing. After deleting stopwords, a term track matrix was created for each dataset based on term frequency (TF) and term frequency–inverse document frequency (TFIDF.) As a result, 10 term track matrices were generated. For the 900 songs, there were 11,602 words from interpretations, and 2,597 words from lyrics.

3.2 Classification and Evaluation Measures

We chose the K-Nearest Neighbors (KNN) classifier since it provides non-linear decision boundaries with relatively simple model parameters to be optimized. We can still get sophisticated enough decision boundaries from the KNN classifier, which in general outperforms naïve Bayes classifiers. In our experiments, we chose seven as the number of neighbors and cosine distance as distance metric since they performed the best in pretests. We performed 10-fold cross validation to evaluate performances of each classifier with different features.

To determine the most useful classification input that provides significantly different results, we applied Friedman's ANOVA test and the Tukey-Kramer "Honestly Significant Difference" (HSD) test. These tests have been commonly used in various MIREX tasks to show whether the significant differences between algorithms' accuracies exist [2]. In our study, each classifier's accuracies per each category were used as inputs for the tests.

4. RESULTS

4.1 Interpretations vs. Lyrics

The accuracies resulting from each of the input conditions (i.e., interpretations, lyrics, and interpretations + lyrics) are better than random, which is 10%. Especially, using interpretations with TFIDF weighting yielded the best performance, and it is statistically different from using lyrics ($p < 0.05$). It turned out that combining the two sources did not improve the classification result. This may be due to the fact that lyrics do not add substantial amount of information to interpretations and yet higher dimensions result in a disadvantage in the classification task. Though the performance differences made by the two weighting methods (e.g., TF vs. TFIDF) were not significant, the results of classifiers using TFIDF always had higher accuracy than TF. Table 1 shows the performance comparison of individual sources and weighting schemes.

Table 1. Classification accuracies of different sources and weighting methods on 900 songs

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

4.2 Accuracy per Category

We also analyzed the confusion matrix (Figure 1) to determine which categories are clearly identified and which are confused when classification was performed based on interpretations. The column of the matrix represents predicted classes, while the row represents the actual ground truth classes. The accuracies of the following six categories were higher than the average: *war*, *religion*, *drugs*, *sex*, *mother* or *father*, and *places*. The four categories, which yielded lower performances than the average, were *old girl/boyfriend*, *loneliness*, *heartache*, and *cheating*. This may partially be due to the correlation between categories. For instance, the fact that *loneliness* and *heartache* share the negative mood may be the reason why *loneliness* is often misclassified as *heartache*.

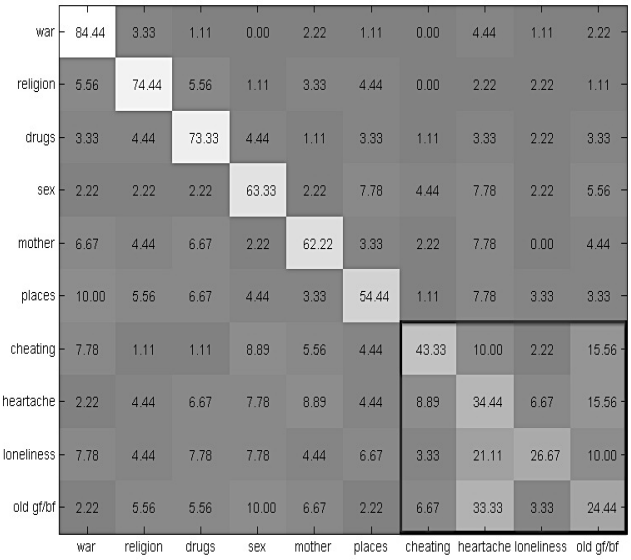

Figure 1. Subject classification confusion matrix using 900 interpretations presented in accuracy rank order

5. CONCLUSIONS AND FUTURE WORK

We have compared two different text sources for classifying songs by subject: user-generated interpretations and lyrics. Our experiment showed that both of them do contain subject-related information to some degree. However, user-generated interpretations outperformed the lyrics by better identifying target categories which rarely overlap with other categories (e.g., *war*, *religion*, *drugs*). Finally, there was no evidence supporting complementary relationships between interpretations and the other two sources.

6. REFERENCES

[1] D. Byrd and T. Crawford, "Problems of music information retrieval in the real world," *Information Processing and Management*, vol. 38, no. 2, pp. 249-272, 2002.

[2] J. S. Downie, "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247-255, 2008

[3] X. Hu, J. S. Downie, K. West, and A. F. Ehmann, "Mining music reviews: promising preliminary results," *In Proc. of 6th Int. Soc. for Music Inform. Retrieval Conf.*, London, UK, Sep. 2005, pp. 536–539.

[4] 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, pp. 287-292.

[5] 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, pp. 441–446.

[4] https://code.google.com/p/stop-words/