Topic Modeling Users' Interpretations of Songs to Inform Subject Access in Music Digital Libraries

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

Jin Ha Lee
Information School
University of Washington
Mary Gates Hall, Suite 370
Seattle, WA 98195
+1.206.685.0153
jinhalee@uw.edu

Craig Willis, J. Stephen Downie
GSLIS
University of Illinois
MC-493, Suite 213
Champaign, IL 61820
+1.217.333.3282
{willis8, jdownie}@illinois.edu

ABSTRACT
The assignment of subject metadata to music is useful for organizing and accessing digital music collections. Since manual subject annotation of large-scale music collections is labor-intensive, automatic methods are preferred. Topic modeling algorithms can be used to automatically identify latent topics from appropriate text sources. Candidate text sources such as song lyrics are often too poetic, resulting in lower-quality topics. Users' interpretations of song lyrics provide an alternative source. In this paper, we propose an automatic topic discovery system from web-mined user-generated interpretations of songs to provide subject access to a music digital library. We also propose and evaluate filtering techniques to identify high-quality topics. In our experiments, we use 24,436 popular songs that exist in both the Million Song Dataset and `songmeanings.com`. Topic models are generated using Latent Dirichlet Allocation (LDA). To evaluate the coherence of learned topics, we calculate the Normalized Pointwise Mutual Information (NPMI) of the top ten words in each topic based on occurrences in Wikipedia. Finally, we evaluate the resulting topics using a subset of 422 songs that have been manually assigned to six subjects. Using this system, 71% of the manually assigned subjects were correctly identified. These results demonstrate that topic modeling of song interpretations is a promising method for subject metadata enrichment in music digital libraries. It also has implications for affording similar access to collections of poetry and fiction.

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

1. INTRODUCTION
The subjects of songs are of great interest to music listeners. Users' strong desire to understand what songs are about is evidenced by millions of online postings discussing and arguing about different interpretations of the meanings of song lyrics. Previous studies also have found that users want subject metadata for music. Bainbridge et al. [1] analyzed 626 music-related online postings from Google Answers, now a defunct Q&A service, and found that "Lyric Story (storyline of song)" was described as one of the information needs. Lee & Downie [2] conducted a large-scale online survey showing that more than 30% of respondents would be likely to use "storyline of music" to navigate music collections, if such an option was available. However, unlike other music metadata such as title, artist, and lyrics, the subject of a song is more difficult to capture. The agent (human or mechanical) needs to comprehend and interpret the lyrics to determine what a song is about. For this reason, enriching a large-scale music digital library (MDL) with additional subject metadata calls for automatic techniques that can efficiently and effectively identify the topics of songs.

Several researchers have attempted to extract subject information from lyrics using both supervised and unsupervised algorithms. For example, Mahedero et al. [3] introduced a naïve Bayes classifier to predict the topic of songs based on lyrics. The algorithm performed well in experiments with 125 songs and five subjects, including "Love", "Violent", "Protest", "Christian", and "Drugs". Similarly, Kleedorfer et al. [4] proposed an automatic subject indexing system that analyzed lyrics using Non-negative Matrix Factorization (NMF), a topic modeling algorithm. Human evaluation of the automatically assigned topics suggests that these unsupervised methods can produce a reasonable number of "good" topics. In addition, Sasaki et al. [5] presented an interface that allows users to navigate music based on topics extracted from lyrics using Latent Dirichlet Allocation (LDA).

While lyrics generally produced positive results, the poetic nature of lyrics can make it difficult for a machine to understand their meaning [6]. Like poetry and some fiction, lyric often use nuanced and deliberately ambiguous language. An auxiliary dataset with elaborated interpretations or explanations of the lyrics may help improve the performance of the automatic systems. In our prior work [6], we demonstrated that interpretations of lyrics are more useful than lyrics themselves when automatically classifying music subjects using supervised methods. In this paper, we shift our focus from supervised methods to an unsupervised algorithm functioning as an automatic music subject discovery system. In the proposed system, a collection of web-mined interpretations of lyrics is used to first identify candidate topics. These topics are then systematically filtered to better represent the subjects of the song lyrics. We propose using prior

JCDL'15, June 21 - 25, 2015, Knoxville, TN, USA

ACM 978-1-4503-3594-2/15/06…$15.00
DOI: http://dx.doi.org/10.1145/2756406.2756936

topic weights and intrinsic topic coherence measurements for this filtering process.

2. THE DATA AND POSSIBLE ISSUES

2.1 Collection

We collected interpretations of lyrics from `songmeanings.com`, where music listeners share and discuss their understanding of lyrics by posting comments about millions of songs. For the experiment, we used songs that appear in both `songmeanings.com` and the Million Song Dataset (MSD)[1], a freely available music collection with a variety of useful audio features and metadata. Of the 58,649 overlapping songs in both collections, we selected 24,436 songs with at least five user interpretations to have sufficient text for our analysis.

In order to conduct an external evaluation of the automatically identified topics, we set aside 422 songs from the test collection with subject labels available from `songfacts.com`. `Songfacts.com` is a website that provides users with a number of browsing options such as subject categories annotated by music experts. From the complete set of 136 subject categories, we selected the six most popular subjects including "war", "parents", "religion", "sex", "drugs", and "heartache". This is consistent with other work [3] in automatic subject identification in music.

2.2 Issues of Lyric Interpretations

Although interpretations often contain subject information, they also contain other types of information that are not useful for browsing music in digital libraries. Examples include general music-related terms, user sentiment, and artist names.

From the preliminary results of topic modeling, we observed several groups of topics that consist of these types of collection-specific terms:

- **General terms**: topics formed around nouns and verbs that universally appear in lyrics interpretations, e.g. *song*, *lyric*, *comment*, *sing*, *music*, *interpret*, *mean*, *understand*, etc.
- **Personal taste/sentiment**: some topics made up of words that express interpreters' personal taste/preference rather than the subject of songs, such as *amaze*, *favorite*, *love*, *awesome*, etc.
- **Music-related terms**: topics mainly related with certain music-related terms, e.g., *play*, *cd*, *rock*, *verse*, etc.
- **Proper nouns**: these topics are composed of names of artists or bands, such as *Bob Dylan*, *Kurt Cobain*, *Oasis*, and *Pearl Jam*.

Some of these issues can be addressed during preprocessing. For example, personal and band names can be identified and removed using standard named-entity recognition techniques. In addition, terms with high document frequencies (DF) can also be handled by establishing a set of corpus-specific stopwords, such as "*lyric*" or "*song*". However, the DF threshold should be chosen carefully, as some important and frequently used subject-related terms might be accidently discarded.

In this work, we address the remaining problems of identifying irrelevant topics in a systematic way. We propose a topic selection technique based on prior topic weights and topic coherence measures, discussed further in section 4.3. We believe that interpretations of poetry and fiction (e.g., novels, short stories) have similar characteristics. While we are limiting ourselves to the lyric case here, we hope to investigate poetry and fiction in future work.

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

3. METHODS

3.1 Topic Modeling

Topic modeling has been widely used to discover latent topics in document collections [7][8][9]. Since our goal is to learn the underlying subjects of lyric interpretations, topic modeling algorithms are well-suited for this work. In general, topic modeling algorithms learn two probability distributions: one over the occurrences of words in each topic, and the other over the contribution of topics to each document. The latter topic distribution allows each document to belong to multiple topics with different probabilities. For the subjects of songs, this is more realistic than clustering techniques that allow only one latent component per item. Latent Dirichlet allocation (LDA) is a powerful generative model that further assumes *a priori* Dirichlet distributions for both the word and topic distributions [7]. In particular, in this work we focus on the latter prior for the topic distribution, which we utilize to judge the popularity of a given topic. Specifically, the Dirichlet parameter, or the prior topic weight, is a vector with elements corresponding to the topics, which tells us the global probability of drawing the multinomial topic distribution. Hence, it eventually determines the contribution of topics to documents: topics with smaller values are rarer in the collection, while at the other extreme can be said to be very popular topics that appear in almost all the documents.

3.2 Evaluation of Topic Modeling

There are two categories of evaluation techniques for topic modeling: intrinsic and extrinsic methods [8]. In this work we use both to measure the quality of the identified topics.

First, we use Pointwise Mutual Information (PMI) as an intrinsic method, as proposed by Newman et al. [8]. Unlike other intrinsic measures, such as perplexity, PMI has been found to be highly correlated with human assessment of topic coherence. This is useful when identifying topics intended for browsing in digital libraries. Newman et al. calculate the PMI of pairs of terms in LDA topic models based on their co-occurrence in Wikipedia.

In this study, we use Normalized PMI (NPMI) as implemented in the Palmetto online tool [9]. NPMI-produced values are bounded between -1 and 1 resulting from the normalization factor, $-\log p(w_i, w_j)$:

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

An NPMI of -1 means the terms never occur together, 0 means independence, and 1 means the terms always occur together. Following Newman et al., we calculate the average NPMI for the top ten terms in each topic and use this as a proxy for topic coherence.

4. EXPERIMENTS

4.1 LDA Implementation

For learning topics, we use the LDA implementation in the MALLET machine learning toolkit [10]. In order to reflect the fact that some topics are more prominent than others in the collection, we conduct hyperparameter optimization. This is also known to result in generating better topics. For this experiment, we selected k=100 topics, as this is closely associated to the

number of subject categories on songfacts.com. Both intrinsic and extrinsic evaluations were performed on the resulting topics.

4.2 Preprocessing

Of the 58,649 songs with records in both songmeanings.com and MSD, only songs with five or more interpretations were used, to ensure sufficient text for topic modeling. This resulted in a collection of 24,437 songs. Only the words consisting of alphabetic letters were considered in order to filter user IDs. General stopwords[2] such as function words were eliminated to increase the quality of learned topics. In addition, terms that appear in more than 40% of the interpretations were also removed. For lemmatization, we used Morphadorner[3], which is also used by Palmetto [9]. In order to remove proper nouns, we used the named-entity recognizer unit in Morphadorner. After preprocessing, a total of 168,846 terms were left.

4.3 Intrinsic Evaluation

To assess the coherence of topics, we calculated the average NPMI of the top ten terms in each topic. As shown in Figure 1, the NPMI values range from -0.19 to 0.26 (mean= 0.07, sd=0.08). Of the 100 LDA topics, 83 have positive NPMI values, suggesting that topics are generally composed of words that are not independent. However, the NPMI values alone are not sufficient for assessing topic quality for use in automatic subject labeling.

In Figure 1, the probability of a topic in the collection is plotted against the associated topic NPMI values, sorted in descending order. A polynomial regression of the NPMI values is also plotted. From this graph, we can see that topics with very high or low probabilities in the collection also have lower NPMI values.

A sample of 12 topics with high (H), medium (M) and low (L) prior topic weights (Dirichlet parameter) are presented in Table 1, along with the associated NPMI values and the top ten topic terms. Topics with higher NPMI values are generally strongly related to potential song subjects. For example, topic M4 (NPMI: 0.24) is clearly about religion or Christianity. However, some topics are clearly outliers. For example, L3 (NPMI: 0.27) is not a subject, but a collection of Spanish terms. Topics with low NPMI values are difficult to associate with a particular subject. In general, they also have lower prior topic weights, such as L2 (NPMI=-0.19).

Most of the topics with very high prior topic weights consist of corpus-specific terms that were described in Section 2.2. The fact that they co-occur more frequently in the collection than the whole Wikipedia may be the reason for their lower NPMI values. Most of the topics with low prior topic weights tend to also have low NPMI values. The three points on the far right side of Figure 1 have very high NPMI values and low prior topic weights. Each of these topics contains non-English words in Spanish, German and French.

In the next section, we demonstrate how the LDA topic probabilities and topic NPMI values can be used to improve subject label quality in an extrinsic evaluation.

4.4 Extrinsic Evaluation

The goal of this evaluation was to determine the effect of different filtering options on the correlation of topics with the ground-truth labels from songfacts.com.

For extrinsic evaluation of the LDA topics, we constructed a ground-truth test set based on songs found in songfacts.com. The resulting collection consists of 422 songs manually labeled with the six most popular subjects. Subjects include "Heartache", "Sex", "Parents", "Religion", "Drugs", and "War". Although the number of songs is not large enough to completely evaluate the quality of all topics, we believe it is sufficient to evaluate topics associated with these six popular subjects.

Instead of manually mapping LDA topics to songfacts.com subjects, we used a majority-voting approach. For each song, we extracted the top three LDA topics. Each songfacts.com category was then mapped to the top-occurring LDA topic based on frequency.

Four different topic-filtering strategies were evaluated based on combinations of NPMI and prior topic weights. First, under the baseline condition all topics were used. In the second case, to filter topics that occur too frequently or infrequently in the collection, only those topics with a collection probability between 0.05 and 0.40 were considered. In the third case, only topics with an NPMI greater than 0.05 were considered. In the fourth and final case, only topics with both a collection probability between 0.05 and 0.40 and NPMI greater than 0.05 are considered.

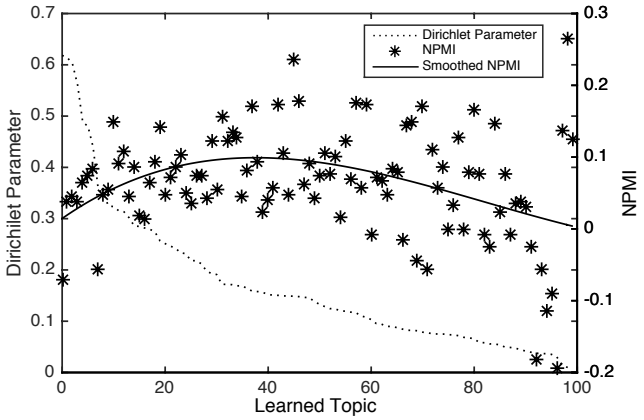

Figure 1. Prior topic weights (Dirichlet parameter) and NPMI values of LDA topics (k=100).

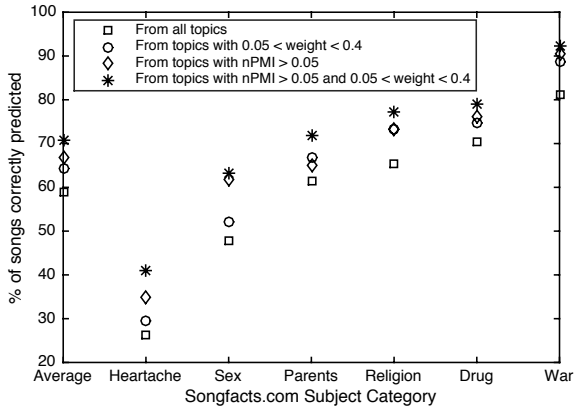

Figure 2. Extrinsic evaluation results (when we assume that top three topics are all equally important).

[2] https://code.google.com/p/stop-words/
[3] http://morphadorner.northwestern.edu/

Table 1. Selected topics from 100 learned topics

Selected Topics	Topic ID	Topic weight	NPMI	Top Words (Top 10)
High Weight	H1	0.62	*-0.07*	awesome, yeah, cd, relate, kinda, lol, total, lot, rock, play
	H2	0.61	0.04	reference, refer, sense, verse, idea, obvious, probable, kind, lot, bit
	H3	0.60	0.05	interpretation, narrator, verse, experience, place, literal, sense, point, speaker, fact
Medium Weight	M1	0.32	0.06	relationship, break, feeling, work, long, leave, girl, hurt, situation, stay
	M2	0.17	0.16	child, father, mother, parent, family, son, dad, brother, daughter, kid
	M3	0.15	0.11	sex, sexual, girl, prostitute, lust, woman, dirty, sexy, whore, desire
	M4	0.15	0.24	god, christian, religion, faith, religious, church, belief, bible, christianity, jesus
	M5	0.13	0.11	drug, addiction, heroin, high, addict, smoke, cocaine, reference, refer, coke
	M6	0.12	0.10	war, fight, soldier, bush, bomb, country, battle, kill, army, military
Low Weight	L1	0.04	*-0.09*	green, river, rise, weezer, edge, lucky, pie, bye, buy, stuff
	L2	0.03	*-0.19*	la, ghost, holly, deaf, vulture, ear, bebot, yeah, gorillaz, bounce
	L3	0.01	0.27	la, spanish, el, en, lo, se, es, mi, una, por

Figure 2 presents the percentage of songs correctly assigned to each of the six popular `songfacts.com` subject categories. The average percentage of correctly labeled songs without filtering is 59%. This increases to 71% with filtering based on a combination of NPMI and prior topic weights. This result suggests both prior topic weights and NPMI are useful criteria when assigning subjects to songs.

In one example, the dominant topic for the category 'Heartache' before filtering is topic H1 (awesome, yeah, cd...). After filtering, the primary topic for category "Heartache" is M1 (relationship, break, feeling...), which is semantically closer to the subject. For the remaining four subject categories, the dominant topic remains the same with or without filtering. "Parents" is mapped to topic M2 (child, father, mother...); "Sex" to M3 (sex, sexual, girl...); "Religion" to M4 (god, Christian, religion...); "Drugs" to M5 (drug, addiction, heroin...) and "War" to M6 (war, fight, soldier...). In all cases, the primary/dominant LDA topics include words that are semantically related to the ground-truth label and also have higher NPMI scores.

These results suggest that the proposed system is effective for detecting song subjects. In addition, the higher average accuracy when either filtering criteria were applied indicates that both criteria are useful for improving subject quality.

5. CONCLUSION

Because lyric terms are so ambiguous, we have presented a method for automatic identification of song subjects based on topic modeling of users' interpretations. We presented techniques for filtering LDA topics using topic coherence values and prior topic weights, and demonstrated how these can be applied to improve the quality of assigned subjects. Intrinsic evaluation using a topic coherence measure has been performed to automate the topic quality assessment process. Extrinsic evaluation using a small ground-truth dataset suggests that this system is effective for automatic subject analysis for possible subject access in MDL. In the future, we plan to expand this work to explore supervised topic modeling using an expanded ground-truth dataset. We also hope to find interpretation sources for poetry and fiction to explore our intuition about providing similar automatically created subject access in digital libraries consisting of these materials.

6. ACKNOWLEDGMENTS

We thank The Andrew Mellon Foundation for their financial support.

7. REFERENCES

[1] 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.

[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] 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.

[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, 287-292.

[5] 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.

[6] K. Choi, J. H. Lee, 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.

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

[8] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," *In Proc. of Human Language Technologies (NAACL-HLT 2010)*, LA, CA, June, 2010, 100-108.

[9] 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.

[10] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.