Semantic Analysis of Song Lyrics

Beth Logan, Andrew Kositsky* and Pedro Moreno
Hewlett Packard Labs
Cambridge, USA.
Beth.Logan@hp.com, Pedro.Moreno@hp.com

Abstract
We explore the use of song lyrics for automatic indexing of music. Using lyrics mined from the Web we apply a standard text processing technique to characterize their semantic content. We then determine artist similarity in this space. We found lyrics can be used to discover natural genre clusters. Experiments on a publicly available set of 399 artists showed that determining artist similarity using lyrics is better than random but inferior to a state-of-the-art acoustic similarity technique. However, the approaches made different errors, suggesting they could be profitably combined.

1. Introduction
The quantity of music available ubiquitously is growing rapidly. There is thus a need for automatic analysis techniques to organize such repositories. Determination of similarity between artists and songs is at the core of such algorithms since it provides a scalable way to index and recommend music.

Many automatic techniques to determine song or artist similarity have been proposed. (e.g. see [1] and references). These approaches are based on analysis of either the acoustic information in the audio or on metadata found on the Web or collected from users. Although some success has been achieved, the systems still fall short of users’ expectations [1]. In this paper we investigate similarity based on a rich source of metadata: lyrics.

Lyrics have several advantages over other forms of metadata. First, the transcriptions of many popular songs are available online. Thus unlike other forms of metadata such as user preferences, lyrics are easy to collect. Also, they are non-subjective; there is only one ‘true’ transcription for a song. This is in contrast to more subjective forms of metadata such as expert opinions or MIDI transcriptions. Finally, lyrics provide a much richer description of the song than simple forms of metadata such as the title, artist and year and arguably contain the true ‘content’ for many songs.

Despite these advantages and their enormous descriptive value, lyrics have received relatively little attention from researchers other than trivially including them as searchable metadata in retrieval systems (e.g. [2]). Scott studied text classification on two tasks containing around 450 folk tunes each but did not compare the results to audio analysis techniques [3]. Brochu and de Freitas developed a framework to jointly model different properties of data and used it to analyze musical scores and associated text annotations, including lyrics, for 100 songs [4]. Although their results are promising the size of their study is too small to draw many conclusions. Additionally, they do not compare their results to audio analysis or investigate very deeply whether the joint model provides more information than one based on just one data source.

In this paper we explore whether the application of text analysis techniques to song lyrics provides meaningful information. Of especial interest is whether such analysis can help determine music similarity and whether it augments other, particularly acoustic, information. We therefore apply a standard semantic text analysis technique to a collection of lyrics to investigate the properties of such data. We explore the use of this analysis to determine artist similarity, comparing the results to a state-of-the-art acoustic similarity technique. For ground truth, we use a large set of user responses collected in a Web survey.

2. Semantic Analysis of Lyrics
Text can be semantically analyzed using various methods. One such technique is Probabilistic Latent Semantic Analysis (PLSA) [5]. PLSA is attractive because it handles both the polysemy and synonymy problems. The approach measures the similarity between text documents by converting each to a characteristic vector. Each component of the vector represents the likelihood that the document is about a pre-learnt topic. Distance in this vector space reflects likely semantic closeness. The topics are characterized by the words that appear frequently in them and are learnt during

* Andrew Kositsky completed an internship at HP Labs Cambridge as part of the Research Science Institute program. http://www.cee.org/rsi/index.shtml
an automatic training process.

3. Preliminary Analysis

In this section we describe our experimental database and explore the feasibility of using PLSA to analyze song lyrics.

3.1. Lyrics Database

As we are not aware of any publicly available popular lyrics database we created our own by mining public websites. Specifically, we collected as many lyrics as we could for the 400 artists covered by the uspop2002 dataset (described in Section 4.1). Most of the data came from azlyrics.com but a number of other sites were mined. We collected lyrics for 15,589 songs in total. This represented 399 of the 400 artists in uspop2002. The remaining artist, Kenny G, is a saxophone player whose songs do not contain lyrics.

3.2. Word Distribution by Genre

Since PLSA models topics using word frequencies, we first study whether the most frequent words in a genre are distinctive. We assign genres to all the artists in the database according to the All Music Guide (www.allmusic.com) and count the most frequently occurring words for the songs in each genre, ignoring a standard list of stopwords. Table 1 shows these results. We see that although Rock and Country have similar vocabularies, other genres such as Newage and Rap are distinctive. We also see that our stopword list would benefit from the addition of ‘lyric-specific’ words such as ‘I’m’ and ‘love’. We shall use this observation later in the paper.

3.3. Topic Models

As described in Section 2 and [5], the first step in PLSA is to learn a set of topics from a text corpus. We used two corpora. The first, denoted NYT, contains documents from two and a half years of the New York Times. The second, denoted LYRICS, is our set of lyrics for the 399 artists augmented with other lyrics from www.azlyrics.com to bring the total number of songs to 41,460. We expect the topics learnt from the latter set to be more suited for the analysis of lyrics.

Table 2 shows the most frequent words for typical topics learnt from the NYT and stemmed LYRICS corpora. We see that PLSA automatically learns prominent topics which are characteristic of the respective data domains. As expected, the NYT models do not contain information that is likely to help differentiate different song or artist styles.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Top 5 words for various topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYT</td>
<td>united government states china american game first team last two</td>
</tr>
<tr>
<td></td>
<td>company business percent companies new people city children dr</td>
</tr>
<tr>
<td></td>
<td>like new ms music film</td>
</tr>
<tr>
<td>LYRICS</td>
<td>don’t feel hate insid god</td>
</tr>
<tr>
<td></td>
<td>love heart feel god sky</td>
</tr>
<tr>
<td></td>
<td>blue beauti sun oh love</td>
</tr>
<tr>
<td></td>
<td>n***** g*** ya f*** yo</td>
</tr>
<tr>
<td></td>
<td>que var de la y</td>
</tr>
</tbody>
</table>

Table 2: The top 5 words for typical topics learnt from the NYT and stemmed LYRICS data. Obscene words have been obscured.

4. Artist Similarity

We now study determining artist similarity using lyrics. Our focus on artist rather than song similarity is driven by the fact that we have a form or ground truth available at the artist level and does not preclude the use of this technique for song-level similarity.

4.1. Ground Truth

We score our results against the publicly available uspop2002 dataset [1]. This set contains acoustic data and semantic metadata for 400 popular artists. These 400 artists are approximately the most popular artists on file sharing and playlist sites in 2002. Consequently, a large proportion of this dataset is from the Rock genre (82%), with Rap (7%), Electronica (4.5%) and Country (3.3%) being the other main categories represented.

We use the so-called “survey” data as the ground truth. In the survey, users were shown an artist randomly chosen from the set of 400 and asked to choose the “most similar” artist from a list of 10 other randomly chosen artists. 10,876 such responses are available in the database (excluding those for Kenny G).

We use two figures of merit to evaluate our automatic similarity techniques as described in [1]. The first is so-called Average Response Rank (ARR). This determines the average rank of the artists chosen from the list of 10 presented in the survey according to the our experimental met-
ric. For example, if our metric agrees perfectly with the human subject, then the ranking of the chosen artist will be 1 in every case, while a random ordering of produces an average ranking of 5.5. In practice, the survey subjects did not always agree so the best possible score is 2.13.

The second figure of merit is First Place Agreement (FPA) which simply counts how many times the similarity metric agrees with the user about the first-place or most similar artist chosen from the list. This metric has the advantage that it is possible to derive significance tests for it and hence evaluate whether variations in values correspond to genuine differences in performance.

4.2. Acoustic Similarity

An advantage of using the uspop2002 dataset is that we can compare our results to previously published similarity techniques. We therefore include results for determining artist/song similarity using a state-of-the-art method based on acoustic information [6]. Briefly, this technique first divides the audio for each artist into frames of 25ms and calculates 20 Mel-frequency cepstral features for each frame. The frames are then clustered using K-means to form a representative signature for each artist. Similarity between artists is calculated using the Earth Mover’s Distance between the representative clusters. In this paper, we study a system with 32 clusters per artist.

4.3. NYT Topic Models

We now calculate artist similarity based on lyrics. We first train topic models as described Section 3.3 and [5]. We then process the lyrics for each artist using each model to form an N-dimensional characteristic vector, where each dimension reflects the likelihood of that artist’s songs being about a pre-learnt topic. To compare artists, we calculate the L1 distance between each vector. We then automatically simulate the user survey by returning the 10 closest artists for each query artist.

Table 3 shows the ARR and FPA for the uspop2002 dataset for artist similarity based on lyrics using various NYT topic models vs acoustic, random and optimal similarity measures. Lower values are better for ARR, higher values are better for FPA.

4.4. Lyric Topic Models

Table 4 shows the ARR and FPA for the uspop2002 dataset for artist similarity based on lyrics using topic models trained on the LYRICS corpus. Again, we also show results for the acoustic technique and random and optimal results.

The first set of results in this table are for models trained using a dictionary of size 91,259. We see that these results are slightly but not significantly better (at least for FPA) than those obtained for the NYT models.

An obvious refinement to text retrieval algorithms is to use stemming. We therefore preprocessed the LYRICS corpus using Porter stemming and built new topic models. This resulted in a dictionary of size 69,566. We then used these new models to recalculate characteristic vectors for each artist and determine artist similarity. The second set of results in Table 4 are for this experiment. We see a small but significant improvement over the non-stemmed results.

Finally, we recall our earlier observation that a number of words are common to many songs and should therefore
Table 5: Analysis of errors comparing accuracy of top rank returned for lyric and acoustic-based similarity measures. Results shown by genre of the target artist presented to the user taking the survey. Results only shown for prevalent genres.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Total Responses</th>
<th>Both Techniques</th>
<th>Lyrics correct</th>
<th>Lyrics incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Incorrect</td>
<td>Acoustic</td>
<td>Acoustic</td>
</tr>
<tr>
<td>Rock</td>
<td>9513</td>
<td>6174 (65%)</td>
<td>1220 (13%)</td>
<td>1627 (17%)</td>
</tr>
<tr>
<td>Rap</td>
<td>316</td>
<td>176 (57%)</td>
<td>48 (15%)</td>
<td>51 (16%)</td>
</tr>
<tr>
<td>Reggae</td>
<td>227</td>
<td>176 (78%)</td>
<td>9 (4%)</td>
<td>39 (17%)</td>
</tr>
<tr>
<td>Country</td>
<td>225</td>
<td>127 (56%)</td>
<td>32 (14%)</td>
<td>44 (20%)</td>
</tr>
<tr>
<td>Latin</td>
<td>201</td>
<td>74 (37%)</td>
<td>48 (24%)</td>
<td>20 (10%)</td>
</tr>
<tr>
<td>Electronica</td>
<td>311</td>
<td>224 (72%)</td>
<td>9 (3%)</td>
<td>72 (23%)</td>
</tr>
</tbody>
</table>

not be included in the topic models since they do not provide any discriminating power. That is, our list of stop words should include additional ‘lyric-specific’ words. To determine these words, we examined the top 1000 words in the 32 topic model built with stemmed data. Any word that occurred in more than half the topics was added to the stop word list, resulting in a reduced dictionary of size 69,303. We then rebuilt the topic models and reran the similarity experiments. The results appear in the third section of Table 4 as “Lyric Iterated”.

These results are significantly better than the previous results and the best results obtained using similarly based on lyrics. They are still worse than the baseline acoustic technique however.

5. Discussion

Although we have demonstrated that determining similarity based on lyrics is feasible, we have also found that it is less useful than acoustic information. Part of the reason for this is that the ground truth that we are using may bias the results toward acoustic similarity. It is likely that people who took the survey thought about the sound of the music rather than the lyrics of the songs. Perhaps then the type of similarity determined by lyrics is irrelevant. Assuming that the similarity measure we are seeking is closely related to that gathered by the survey, we now investigate whether lyrics bring anything of use.

In Table 5 we show an analysis of the errors in common and unique to the lyric and acoustic similarity distances. The results shown are for the best performing lyric similarity measure. The errors are grouped by genre of the target artist presented to the user taking the survey.

From this table we see that the lyrics similarity technique is intrinsically better at determining similar songs in the Latin category and has trouble with Electronica and Reggae. Conversely the acoustic technique is intrinsically better at Electronica and Country and poor at Latin. This is because Latin music is typically not in English. Therefore, users are less likely to mentally group Latin songs with say Rock songs, even if they are acoustically similar. Conversely, Electronica tends to have fewer lyrics than other songs on average so it is no surprise that the acoustic similarity technique gives superior results for that category. Both techniques have trouble with Reggae music.

This analysis indicates that a combination of these measures could result in an improved similarity measure. Determination of an optimal combination strategy is the subject of future work.

6. Conclusions

We have explored a technique to automatically analyze song lyrics and determine artist similarity based on this. We evaluated the technique on a publicly available dataset of 399 popular artists, comparing it to an acoustic similarity technique. For ground truth, we used data collected in a user survey. Similarity based on lyrics was found to be better than random but inferior to acoustic similarity, at least for the ground truth used. However, the errors made by each technique were not randomly distributed suggesting that the best technique would be a combination of both.

7 References