Building and Combining Document and Music Spaces for Music Query-By-Webpage System

Ryoei Takahashi[1], *Yasunori Ohishi*[1], *Norihide Kitaoka*[1], *and Kazuya Takeda*[1]

[1]Graduate School of Information Science, Nagoya University

{takahasi,ohishi}@sp.m.is.nagoya-u.ac.jp, {kitaoka,kazuya.takeda}@nagoya-u.jp

Abstract

Building and combining document and music spaces of songs are discussed for a new music recommendation application, which uses commonly read texts such as Web log as query input. The most important application of this flexible recommendation system is its music query-by-Webpage, from which a song that appropriately matches Webpage is automatically played. The key idea of the proposed system is to train a linear transformation between document and music spaces so that query documents can be mapped onto a music space in which similarities based on acoustic characteristics is represented.

The basic system has been trained using 2,650 pairs of song and review texts. Through experimental evaluations, we show the effectiveness of the system, which is three times better than the previous system. Web text as a training corpus and a bigram representation for the document vector are also investigated for the purpose of improving the system, and their effectiveness is also confirmed.

Index Terms: Music, information retrieval, music similarity, latent semantic analysis, multimedia databases

1. Introduction

Building a similarity measure between songs is one of the most fundamental issues in music information retrieval systems. Various similarities have been developed for a wide range of music retrieval tasks [1, 2]. In most query-by-keyword systems, text similarities defined by titles, artist names, and lyrics, are commonly used [3]. Acoustic similarities between songs are mainly studied for song classification tasks, such as genre recognition [4].

In addition, few have tried enumerating similarities between words and songs for music query-by-text retrieval and music annotation systems. Kumamoto et al. [5] associated words expressing impressions, such as "beautiful" and "sad", with particular acoustic characteristics of music through multiple-regression analysis. Turnbull et al. [6] manually selected musically informative words and associated them with acoustic cues using a stochastic approach, i.e., the supervised multiclass naive Bayes model. Using this model, they evaluated retrieval performance on the basis of a single word query. They also evaluated annotation performance by tagging a song with a set of relevant words. Whitman et al. [7] used a classifier that discriminates relevant and irrelevant words to a song for the purpose of automatically generating song reviews. Slaney [8] tried to connect words with environmental sounds.

In these works, individual correspondences between songs and words were modeled,; however, relationships among songs and words were not explicitly modeled.

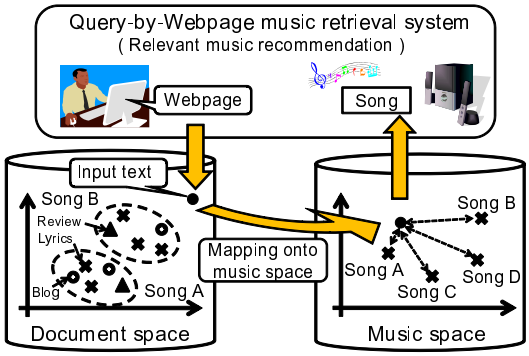

Figure 1: Query-by-Webpage music retrieval system that bridges document and music spaces.

In this paper, we characterize a song by two different vector spaces, the document space and the music space. The document space is a latent semantic space defined by the documents related to the song, such as review texts, lyrics, and Web texts describing the song. The music space is a signal acoustic space defined by the spectral/temporal characteristics of a song. By combining document and music spaces we can build a vector space in which "closeness" among songs and texts is preserved. We can then retrieve songs that are located near the input text in the document space, and we can also retrieve songs that are musically similar to those songs. The idea behind our system is depicted in Figure 1. The document space $\{\mathbf{d}\}$ and music space $\{\mathbf{a}\}$ are associated by the linear transformation $\mathbf{a} = W\mathbf{d}$. Through this transformation, we can retrieve songs that are closest to the mapped document vector in the music space, even when the song has not been used in building the document space.

In Section 2 of this paper, we describe how document and music spaces are built and combined. In Section 3, we discuss the implementation and evaluation of the basic system. In Section 4, we discuss investigations into further improving the document space, and in Section 5 we conclude the paper.

2. Basic Algorithm

To measure the similarity between a song and related texts, we first build two linear vector spaces of text documents and music signals. We then find a linear transformation between two

spaces. In this section, we discuss our method for building document and music spaces as well as a method for combining the two spaces.

2.1. Building document space

Latent semantic analysis [9] is used as a basic method for building the document space of songs. To emphasize words that are relevant to song information, we use term frequency-inverse document frequency (TF-IDF) for document representation in stead of simple word frequency. Given J sets of texts related to J songs, the $I \times J$ matrix of TF-IDF weight $\mathbf{X}$ is calculated.

$$X_{ij} = \frac{tf_{ij}}{\sum_i tf_{ij}} \times \log \frac{J}{df_i}, \tag{1}$$

where I is the number of words that span the document space, and tf and df are term frequency and document frequency, respectively. The $(i,j)^{th}$ element of X is therefore the (relative) frequency of the i^{th} word in the documents corresponding to the j^{th} song, i.e., the j^{th} column vector of X represents document vector $\mathbf{x}_j$ of the j^{th} song.

The document space $\{\mathbf{d}\}$ is obtained by the singular value decomposition of X:

$$X = USV^{\mathrm{T}}. \tag{2}$$

Assuming that the diagonal elements of S, i.e., the singular values of X, are arranged in descending order, the dimensions of the document vector are thus reduced to arbitrary number N $(N < I)$ by truncating the higher order element as follows:

$$\mathbf{d} = U_N^{\mathrm{T}}\mathbf{x}, \tag{3}$$

where U_N represents the 1^{th} to N^{th} columns of matrix U.

2.2. Building acoustic space

The following five acoustic measurements, based on short-time spectral information, are used for building the music space [10].

2.2.1. Spectral centroid

The spectral centroid is defined as the center of gravity of the spectrum:

$$C_t = \frac{\sum_{k=1}^{K} M_t[k] \times k}{\sum_{k=1}^{K} M_t[k]}, \tag{4}$$

where $M_t[k]$ is the magnitude of the Fourier transform at frame t and frequency bin k. The centroid is a measure of spectral shape, and higher centroid values correspond to 'bright' textures with higher frequencies.

2.2.2. Spectral rolloff

Spectral rolloff is defined as the frequency R_t below which 85% of the spectrum distribution is concentrated:

$$\sum_{k=1}^{R_t} M_t[k] = 0.85 \times \sum_{k=1}^{K} M_t[k]. \tag{5}$$

Rolloff is another measure of spectral shape.

2.2.3. Spectral flux

Spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions:

$$F_t = \sum_{k=1}^{K} (N_t[k] - N_{t-1}[k])^2, \tag{6}$$

where $N_t[k]$ and $N_{t-1}[k]$ are respectively the normalized spectra at current frame t and previous frame $t-1$. Spectral flux measures the amount of local spectral change.

2.2.4. Spectral flatness

Spectral flatness measure (SFM) is computed by the ratio of the geometric mean to the arithmetic mean of the power spectrum value:

$$SFM_t[i] = \frac{\left\{\prod_{k \in B_i} P_t[k]\right\}^{1/K_{B_i}}}{\frac{1}{K_{B_i}} \sum_{k \in B_i} P_t[k]}, \tag{7}$$

where $P_t[k]$ is the power spectrum at frame t, and K_{B_i} is the total frequency bin number in frequency band B_i. The width of frequency band B_i is a 1/4 octave, and there are 24 bands between 250 Hz and 16 kHz.

2.2.5. Zero-crossing rate

The frequency at which the time domain signal crosses zero is calculated as:

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|, \tag{8}$$

where the function $sign$ is 1 for positive arguments and 0 for negative arguments, and $x[n]$ is the time domain signal for frame t. Time domain zero-crossing measures noisiness of the signal.

These features are calculated by Marsyas ver. 0.2.10 [11]. After calculating the features over a sequence of frames, the dynamic features are derived by calculating the regression coefficient of each element over consecutive frames. Using these $L = 10$ dimensional feature vectors of the training data, we create a vector quantization (VQ) codebook of which the size is M. Then, a normalized code histogram of the VQ results of the song is used as M dimensional acoustic feature vector $\mathbf{a}$, representing the musical characteristics of the song.

2.3. Associating two spaces

Document vector $\mathbf{d}$ and acoustic vector $\mathbf{a}$ are associated through linear transformation:

$$\hat{\mathbf{a}} = W\mathbf{d}. \tag{9}$$

The $M \times N$ transformation matrix W can be trained using pairs of document and acoustic vectors $\{(\mathbf{d}_j, \mathbf{a}_j)\}_{j=1,2,\cdots}$ with the minimum squared error criterion:

$$\hat{W} = \underset{W}{\operatorname{argmin}} \sum_j ||\mathbf{a}_j - W\mathbf{d}_j||^2. \tag{10}$$

Matrix $\hat{W}$ that minimizes the squared error is determined by:

$$\hat{W} = \left(\left(\sum_j \mathbf{d}\mathbf{d}'\right)^{-1} \sum_j \mathbf{d}\mathbf{a}'\right)^{-1}, \tag{11}$$

where $\mathbf{x}'$ represents the transposition of vector $\mathbf{x}$.

Table 1: Performance of the baseline system.

proposed method	(open)	MRR = 0.210
	(open)	mAP = 0.351
	(close)	mAP = 0.816
naive Bayse [6]	(open)	mAP = 0.109

3. Evaluation of Baseline System

The basic performance of the system was evaluated through the following experiments.

3.1. Experimental setup

A prototype system was implemented with 2,650 pop songs to evaluate the feasibility of the proposed method.

Thirty-second previews, as well as the reviews of 2,650 pop songs, were extracted from the music download site Mora [12]. The average length of each review is 2.74 sentences, and in total, 11,428 different words were included after morphological analysis by Chasen ver. 2.3.3 [13]. Only nouns, adjectives, and verbs were used for building the document space; therefore, dimension I of the document vector was 10,578.

For acoustic analyses, a 32-ms analysis window was applied every 16 ms, and the acoustic feature vector, described in Section 2.2, was calculated at each frame. The sampling rate of the song signal was 16 kHz, and the SFM order was 14.

Since the dimension of the document space was further reduced through SVD, the size of the transformation matrix between document and music spaces, i.e., W, is 1,024 (the dimension of acoustic vector M) $\times$ 1,024 (the truncated dimension of document N).

To evaluate the system under the 'open' condition, 2,650 song and review pairs were divided into five sets. Four of them were used for training, and the other one was used for tests. The document vector of a review text was calculated and then mapped onto an acoustic space through transformation. Then the rank-ordered song list based on the distance from the mapped vector of the review text was generated as an output list of the query.

As for the performance evaluation measure, the mean reciprocal rank (MRR) given by:

$$\mathrm{MRR} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{r_k} \tag{12}$$

was used, where N is the total number of test samples and r_k is the rank order of the k^{th} song for which the review text was given.

3.2. Evaluation results

The results of the evaluation are listed in Table 1. As shown in the table, a MRR of 0.21 was obtained using the above mentioned experimental conditions. To compare this result with that of the previously proposed music query-by-text system on the basis of the naive Bayes model [6], we also calculated the mean average precision (mAP) of our results. The performance of the proposed system, with a mAP of 0.351, was approximately three times better than the query-by-text system, with a mAP of 0.109 [6]. These results thus clarified the effectiveness of combining document and music spaces.

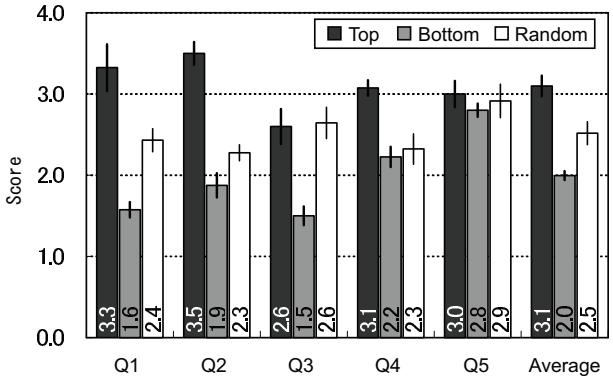

Figure 2: Subjective evaluation results of the three sets of query results, i.e., top 10 songs (Top), bottm 10 songs (Bottom) and randomly selected 10 songs (Random). The five query sentences are:
Q1. An exhilarating song like a summer breeze that presages the first step in a new beginning. Q2. A sensitive ballad that conjures up sentimental thoughts. Q3. An up tempo rock tune filled with guitar riffs and pop sense. Q4. Music that soothes the tired soul. Q5. A rebel yell song with a pulsating bottom beat that-in a time of difficulty-sends a resolute message to everyone.

The system was also evaluated through a subjective test. In this test, four subjects evaluated three sets of songs to determine how appropriately the set of songs corresponds to the input query sentences on a scale of 0 to 4. The three sets of songs consist of 1) top 10 ranked songs; 2) bottom 10 ranked songs; and 3) 10 randomly selected songs. As shown in Figure 2, the mean scores of the top 10 ranked songs are consistently higher than those of the randomly selected songs,; therefore, the effectiveness of the system was also confirmed from a subjective viewpoint.

4. Improving Document Space

To further improve the baseline system, the training corpus and the representation of the document vector were investigated for the purpose of building a better document space.

4.1. Using Web text for training

Since each review describes a particular song, training the system using song and review pairs is reasonable. However, review text is not always available for all songs, and a system that is trained by the review texts may not work well for general texts, such as Web logs. We have therefore trained the document space using Web texts that are collected from the top 100 Web pages of Google search results for a song title and artist's name as query key words. The size of the training corpora for the document spaces of reviews (baseline system) and Web pages are listed in Table 2. Approximately, a 20-times larger corpus than that of the baseline system is used.

4.2. Using word bigrams

A TF-IDF matrix represents word concurrence information; however, word sequence information is not considered. Therefore, instead of using the TF-IDF matrix, the word bigram is

Table 2: Size of the training corpora.

		review articles	web pages
vocaburary	noun	8,709	15,728
	adjective	248	364
	verb	1,611	2,328
word count	n.+adj.+v.	116,153	2,345,919

Table 3: Performance of the Improved Systems.

document vector	training corpus	MRR
tf-idf	review texts	0.210
tf-idf	web texts	0.739
bigram	review texts	0.312
bigram	web texts	0.794

tested while taking into account the word sequence information in the document space. Word bigrams (20,961) that occur more than once are used for building the document space of the song review, and word bigrams (27,448) that occur more than nine times are used for the document space of the Web texts.

4.3. Results

Results related to the above improvements are summarized in Table 3. The MRR scores for the Web texts were much higher than those for the review texts. This is because the size of the training document for each song is larger in Web documents than in review documents. Note that input query texts are also taken from search result pages, and therefore the average length of the sentences is about 10 times longer than that of the review texts.

In both training documents, bigrams performed better than TF-IDF and finally a MRR of 0.793 for the Web text query was achieved. With these results, we clarified the feasibility of the music recommendation system using commonly found Web texts as input.

5. Conclusion

We combined document and music spaces to build a vector space on which "closeness" among songs and texts can be defined. A music query-by-Webpage system was implemented on the basis of the combined vector space.

In an experimental evaluation using 2,650 song and review document pairs, we evaluated the performance of the system in terms of MRR and mAP. Experimental results confirmed the proposed system is effective, having a mAP that is three times higher than that of the previous system (0.351 to 0.109). Further improvements on the use of Web texts as a training corpus and the use of bigrams as a document representation were also investigated. With both approaches, performance improved, and finally, an MRR of 0.793 was achieved.

A lot of future work is needed to refine this system; however, scaling the system is now a key issue. Particularly needed is development of the method that use very large Web documents for training the higher order n-gram, which we expect will bolster the performance of the proposed system.

6. References

[1] J. J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," *Proceedings of the ISMIR*, 2002.

[2] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measures," *Proceedings of the ISMIR*, 2003.

[3] P. Knees, E. Pampalk, and G. Widmer, "Artist classification with web-based data," *Proceedings of the ISMIR*, 2004.

[4] E. Pampalk, A Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," *Proceedings of the ISMIR*, 2005.

[5] T. Kumamoto and K. Ohta, "Design, implementation, and opening to the public of an impression based music retrieval system," *Trans. Japanese Society for Artificial Intelligence*, vol. 21, no. 3, pp. 310–318, 2006 (in Japanese).

[6] D. Turnbull, L. Barrington, and G. Lanckriet, "Modelling music and words using a multi-class naive bayes approach," *Proceedings of the International Symposium on Music Information Retrieval*, 2006.

[7] B. Whitman and D. Ellis, "Automatic record reviews," *Proceedings of the International Symposium on Music Information Retrieval*, 2004.

[8] M. Slaney, "Semantic-audio retrieval," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002.

[9] S. Deerweester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 1, pp. 391–407, 1990.

[10] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[11] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organized Sound*, vol. 4, no. 3, pp. 169–175, 2000.

[12] Mora, "http://mora.jp/," .

[13] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, "Morphological analysis system chasen 2.3.3," *Nara Institute of Science and Technology*, 2003.