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

Ryoei Takahashi, Yasunori Ohishi
Norihide Kitaoka, Kazuya Takeda

Graduate School of Information Science, Nagoya University, Japan
New music retrieval system

- Music Query-By-Webpage System
  - Songs that appropriately match Webpage automatically selected
New music retrieval system

- Music Query-By-Webpage System

Songs that appropriately match Webpage automatically selected

- Text analysis
- Ranking

Play list

- Meja
- The Corrs
- Nick Lowe
- Faye Wong
- Tom Jones

- 「Rainbow」
- 「Breathless」
- 「Cruel to be kind」
- 「Dreams」
- 「It’s not unusual」
New music retrieval system

- **Music Query-By-Webpage System**

  - Songs that appropriately match Webpage automatically selected
  - Text analysis
  - Ranking

**Play list**

- Meja 「Rainbow」
- The Corrs 「Breathless」

Songs associated with words in documents (Webpages)
Previous work

Building similarity measures between songs

Query-by-keyword systems
- Word similarities defined by titles, artist names ...

Song classification tasks
- Acoustic similarities between songs

Music annotation systems
- Words expressing impressions and acoustic cues [Kumamoto et al.]
- Musically informative words and acoustic cues [Whitman et al., Turnbull et al.]

But

Individual correspondences between songs and words
Proposed method

- Two different vector spaces

Document space $\{d\}$ and music space $\{a\}$ associated by linear transformation $a = Wd$
Basic algorithm (1)

- **Building document space \( \{d\} \)**
  - Term frequency-inverse document frequency (TF-IDF)
  - TF-IDF weight matrix \( X (I \times J) \) \((X \equiv [x_1, \ldots, x_j, \ldots, x_J])\)
    \[
    X_{i,j} = \frac{t f_{i,j}}{\sum_j t f_{i,j}} \times \log \frac{J}{d f_i}
    \]
    - \( t f \): Term (word) frequency
    - \( d f \): Document frequency
  - \( I \): Number of words, \( J \): Numbers of songs

**Document for Song \( j \)**

- Up-tempo rock tune filled with riffs and pop sensibility

- **Document vector** \( x'_j \)
  - “tempo” “rock” “techno” ... “guitar” “riff”
  - \([0.08 \ 0.3 \ 0 \ \ldots \ 0.7 \ 0.2]\)

- **Singular value decomposition**
  \[
  X = U S V^T
  \]

- Reduced dimensions of document vector \( X \)
  \[
  d = U_N^T x \quad U_N : 1^{\text{st}} \text{ to } N^{\text{th}} \text{ columns of } U
  \]
Basic algorithm (2)

- Building music space \( \{a\} \)
  - Acoustic characteristics
    - Spectral centroid, rolloff, flux
    - Zero-crossing rate
    - Energy, rhythm
  - Vector quantization (VQ) codebook (Size \( M \))
    - Feature vectors of training data
    - LBG Algorithm
  - Normalized code histogram of VQ results
    - \( M \)-dimensional acoustic feature vector \( a \)

- Song \( j \)
  - Sequence of acoustic features
- Codebook
- Clustering
- Spectral centroid
- Spectral rolloff
- Spectral flux
- 0.85
Basic algorithm (3)

- Associating document and acoustic vectors through linear transformation
  \[ \hat{a} = Wd \]

  - Transformation matrix \( W (M \times N) \) trained using pairs of document and acoustic vectors
    \[ \{ (d_j, a_j) \}_{j=1,2,...} \]
  - Minimum squared error criterion
    \[ \hat{W} = \arg\min_W \sum_j \| a_j - Wd_j \|^2 \]
    \[ \hat{W} = \left( \left( \sum_j dd^T \right)^{-1} \sum_j da^T \right)^{-1} \]
Evaluation of baseline system

- Implemented with 2,650 pop songs
  - Japanese music download site: “Mora”
    - 30 s previews and reviews
  - Document space
    - Morphological analysis by Chasen ver. 2.3.3
    - Nouns, Adjectives and Verbs (I : 10,578)
  - Music space
    - Applied 32-ms analysis window every 16 ms
    - Calculated acoustic feature vector at each frame
- Size of transformation matrix $W$
  - $1,024$ (Dimension $M$ of acoustic vector)
  - $\times 1,024$ (Truncated dimension $N$ of document)
Experimental setup

- **Evaluation under ‘open’ condition**
  - Divided 2,650 song and review pairs into five sets

- **Evaluation measure**
  - Mapped document vector $\mathbf{d}$ of query onto acoustic space $\mathbf{\hat{a}}$ through transformation $W$
  - Generated rank-ordered song list based on distance between $\mathbf{\hat{a}}$ and $\{\mathbf{a}_j\}_{j=1,2,...}$
  - Mean reciprocal rank (MRR)

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

- $N$: number of test samples
- $r_k$: rank order of $k^{th}$ song for which review was given
Evaluation results

- MRR of 0.21 obtained
- Comparison with previous results
  - Music query-by-text system based on Naive Bayes [Turnbull et al.]
  - 3 times better than previous system, with mean average precision (mAP)

<table>
<thead>
<tr>
<th>Proposed method</th>
<th>(open)</th>
<th>MRR = 0.210</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(open)</td>
<td>mAP = 0.351</td>
</tr>
<tr>
<td></td>
<td>(closed)</td>
<td>mAP = 0.816</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>(open)</td>
<td>mAP = 0.109</td>
</tr>
</tbody>
</table>

Results clarified effectiveness of combining document and music spaces
Improving document space

- Using Web text for training
  - Trained document space using Web texts collected from **top 100 Web pages** of Google search results for song title and artist name as query key words

- Using word bigrams
  - Trained document space using word bigrams while considering **word sequence information**

<table>
<thead>
<tr>
<th>Document vector</th>
<th>Training corpus</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>Review texts</td>
<td>0.210</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Web texts</td>
<td>0.739</td>
</tr>
<tr>
<td>Bigrams</td>
<td>Review texts</td>
<td>0.312</td>
</tr>
<tr>
<td>Bigrams</td>
<td>Web texts</td>
<td>0.794</td>
</tr>
</tbody>
</table>
Improving document space

- Using Web text for training
  - Trained document space using Web texts collected from **top 100 Web pages** of Google search results for song title and artist name as query key words

- Using word bigrams
  - Trained document space using word bigrams while considering **word sequence information**

<table>
<thead>
<tr>
<th>Document vector</th>
<th>Training corpus</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>Review texts</td>
<td>0.210</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>Web texts</td>
<td>0.739</td>
</tr>
<tr>
<td>Bigrams</td>
<td>Review texts</td>
<td></td>
</tr>
<tr>
<td>Bigrams</td>
<td>Web texts</td>
<td>0.794</td>
</tr>
</tbody>
</table>

Feasibility of music recommendation system using arbitrary Web texts as input
Subjective test

- Four subjects evaluated three sets of songs to determine how appropriately they corresponded to input query sentences on a scale of 0 to 4
  1. Top 10 ranked songs
  2. Bottom 10 ranked songs
  3. 10 randomly selected songs

![Bar chart showing score comparison between Top, Bottom, and Random sets for Q1 to Q5 and Average.]

- **Top**: Black bars
- **Bottom**: Dark gray bars
- **Random**: Light gray bars

*Example scores:*
- Q1: Average Top 3.5, Bottom 2.5, Random 2.0
- Q2: Average Top 3.8, Bottom 2.8, Random 2.3
- Q3: Average Top 3.2, Bottom 2.2, Random 1.8
- Q4: Average Top 4.0, Bottom 3.0, Random 2.5
- Q5: Average Top 4.2, Bottom 3.2, Random 2.7

*Average scores for all questions:*
- Top: 3.7
- Bottom: 2.7
- Random: 2.3
- Overall Average: 3.1
Subjective test

Four subjects evaluated three sets of songs to determine how appropriately they corresponded to input query sentences on a scale of 0 to 4:

1. Top 10 ranked songs
2. Bottom 10 ranked songs
3. 10 randomly selected songs

“Sensitive ballad that conjures sentimental thoughts”
Subjective test

- Four subjects evaluated three sets of songs to determine how appropriately they corresponded to input query sentences on a scale of 0 to 4:
  - (1) Top 10 ranked songs
  - (2) Bottom 10 ranked songs
  - (3) 10 randomly selected songs

"Sensitive ballad that conjures sentimental thoughts"
Let’s listen

- Demo system “text 2 music”

“Up-tempo rock tune filled with guitar riffs and pop sensibility”
Conclusion

- Built document and music spaces on which “closeness” among songs and texts can be defined and combined
- Implemented music query-by-Webpage system based on combined vector space space
  - Proposed system was effective, having a mAP three times higher than previous system
  - Improving document space
    - Use of Web texts as training corpus
    - Use of bigrams as document representation
- Future work
  - Use very large Web documents for training higher order n-grams
Thank you!

Yasunori Ohishi
ohishi@sp.m.is.nagoya-u.ac.jp