楽曲のレビューと音響特徴量との関連付けの検討 *

◎高橋量衛, 大石康智, 北岡教英, 武田一哉 (名大), 伊藤克亘 (法政大)

1 はじめに

インターネットを介して大規模楽曲データベースにアクセスし，ユーザが，何千，何万という大量の楽曲を所有することができるようになった．今後は，これらの大量の楽曲をユーザが管理し，効率よく検索して鑑賞するための技術が必要となる．現在の標準的な楽曲検索システムでは，好きな曲を検索するために，ユーザ自身が適切な検索クエリを入力しなければならない．

本研究では，ユーザが作成した文章，ユーザが眺めている Web ページといったテキストデータを入力したときに，複数の語彙の共起関係から楽曲を検索することのできるシステムを提案する．例えば，図 1 のように，ユーザが閲覧している Web ページをテキスト解析することによって，そのページを表現するにふさわしい BGM を流すシステムである．そのためには，語彙の共起関係が表現される空間と，楽曲の音響的特徴空間との関連付けを行う必要がある．これまで,「明るい」や「静かな」のような印象語を検索クエリとするシステム [1] は，すでに提案されているが，印象語と楽曲との関連付けは聴取実験に基づくものであり，音響的特徴空間との関連付けに関しては検討されていない．また，本研究では印象語に限らず，テキストデータに出現するあらゆる語彙の共起関係に着目する．これにより，楽曲を入力したときに，音響的特徴空間と語彙空間との関連付けから，楽曲を解説することのできる文章 (レビュー) を自動生成するという応用例も考えられる．

本報告では，初期実験として，楽曲を解説したレビューと，楽曲の音響的特徴との関連付けを試みる．レビューを表現するための文書ベクトルと楽曲の音響的特徴を表現するための音響ベクトルを提案し，これらを線形変換によって関連付けることを考える．

2 楽曲のレビューと音響特徴量との関連付け手法

楽曲のレビューを表現するための文書ベクトルと，楽曲の音響的特徴を表現するための音響ベクトルについて述べる．さらにこれら 2 つの特徴ベクトルを線形変換によって関連付けるための変換行列の推定手法について述べる．

2.1 TF-IDF を利用した文書ベクトルの抽出

楽曲 j を解説したレビューを多次元のベクトル $\mathbf{x}_j$ で表現し，文書ベクトルと呼ぶ．この文書ベクトル $\mathbf{x}_j$ の i 次元目の要素 $x_{i,j}$ は，形態素 t_i に関して以下の式で計算される TF-IDF(term frequency - inverse document frequency) による重みとする．

$$x_{i,j} = \frac{tf_{i,j}}{\sum_i tf_{i,j}} \times \log \frac{J}{df_i} \tag{1}$$

ここで，楽曲 j のレビューにおける形態素 t_i の出現頻度を $tf_{i,j}$，すべての楽曲のレビューのうち，形態素 t_i を含むレビュー数を df_i，楽曲の総数 (レビューの総数) を J とする．レビューの集合を行列 $X = (\mathbf{x}_1, \cdots, \mathbf{x}_j \cdots, \mathbf{x}_J)$ と記述する．X は $I \times J$ の行列であり，I は考慮する形態素の総数である．曲数が増える（レビューの数が増える）につれて，形態素の総数 I も増加する．しかし，1 つのレビューに出現する形態素は限られるため，行列 X は 0 の要素が多いスパースで高次元の行列となる．そこで，以下のように行列 X の特異値分解 [2] を行う．

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

*Association between music review and acoustic features. by R. Takahasi, Y. Ohishi, N. Kitaoka, K. Takeda (Nagoya Univ.), K. Itou (Hosei Univ.)

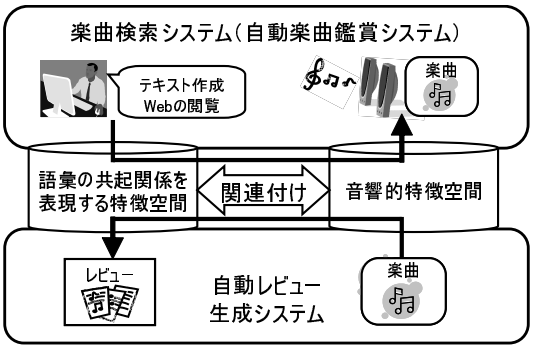

図 1: テキストデータと音響的特徴との関連付けを応用したシステムの例

ここで，S は $J \times J$ の非負要素の対角行列であり，対角要素は絶対値の降順に並んでいるものとする．直交行列 U のうち，絶対値の大きな特異値に対応する第 1 列から第 k 列を取り出した行列 U_k を用いて，I 次元の文書ベクトル $\mathbf{x}_j$ を以下のように k 次元に削減することができる．

$$\mathbf{t}_j = U_k^T \mathbf{x}_j \tag{3}$$

次元削減した $\mathbf{t}_j$ を楽曲 j のレビューを表現する文書ベクトルとして利用する．

2.2 クロマベクトルを利用した楽曲の音響的特徴抽出

音響的特徴量として，標本化周波数 16kHz の音響信号のパワースペクトル $P(f,t)$ (時刻 t，対数スケール周波数 f，STFT 窓幅 $256ms$，フレームシフト $80ms$) から，12 次元のクロマベクトル $\mathbf{v}(t)$ を求める [4]．$\mathbf{v}(t)$ の各次元 $v_c(t)$ は，12 音名の各音名 $c(c = 1, 2, \ldots, 12)$ の周波数のパワーを複数のオクターブ h に渡って加算したもので，

$$v_c(t) = \sum_{h=Oct_L}^{Oct_H} \int_{-\infty}^{\infty} BPF_{c,h}(f) P(f,t) df \tag{4}$$

と定義する．$BPF_{c,h}(f)$ は，音名 c，オクターブ h の位置のパワーを通過させるバンドパスフィルタで，$Oct_L = 3$ から $Oct_H = 8$ まで，130Hz〜7.9kHz の 6 オクターブに渡るように設定した．

また，クロマベクトル $\mathbf{v}(t)$ の各要素の前後 2 点の計 5 点に渡って直線回帰することによって得られる回帰係数を動的特徴量 $\Delta\mathbf{v}(t)$ とする．したがって，各時刻ごとにクロマベクトルとその動的特徴量が計算される．全楽曲から求めた特徴ベクトル (クロマベクトルとその動的特徴量) の集合を N 個のクラスにベクトル量子化し，各セントロイドを表すコードブックを求める．次に各楽曲ごとに，特徴ベクトルの集合をコードブックに基づいてクラスタリングし，その頻度分布を楽曲の音響的特徴を表現する音響ベクトルとして利用する (楽曲 j の音響ベクトルは $\mathbf{a}_j$ と表す．$\mathbf{a}_j$ の要素数は，ベクトル量子化におけるクラス数 N である)．

2.3 変換行列の推定

文書ベクトル $\mathbf{t}_j$ と音響ベクトル $\mathbf{a}_j$ を以下のような線形変換によって関連付けることを考える．

$$\mathbf{a}_j = W\mathbf{t}_j \tag{5}$$

ここで変換行列 W は，音響ベクトル $\mathbf{a}_j$ と W に文書ベクトル $\mathbf{t}_j$ をかけた $W\mathbf{t}_j$ との 2 乗誤差 $||\mathbf{a}_j - W\mathbf{t}_j||^2$ が J 曲すべてに関して最小となるように求められる．

$$\hat{W} = \underset{W}{\operatorname{argmin}} \frac{1}{J} \sum_{j=1}^{J} ||\mathbf{a}_j - W\mathbf{t}_j||^2 \tag{6}$$

ここで推定する変換行列 W は正方行列とした．すなわち，文書ベクトル $\mathbf{t}_j$ の要素数 k と音響ベクトル $\mathbf{a}_j$ の要素数 N は等しい．

3 評価実験

3.1 使用データ

音楽ダウンロードサイト Mora[3] における試聴曲 (約 30 秒程度) と，その曲を解説したレビューを提案手法の学習と評価に利用する．試聴曲とレビューは 1 対 1 の関係にあり，アルバム曲全体を解説したレビューは，今回使用しない．その結果，全部で 2,705 曲の音響信号とレビューを集めることができた．

レビューあたりの平均文章数は，2.74 文であった．茶筅 ver.2.3.3 を利用して形態素解析を行った結果，形態素の種類は 11,250 であった．そのうち品詞を名詞，動詞，形容詞に限定した場合，形態素の種類は 10,462 であり，これを文書ベクトル $\mathbf{x}_j$ の要素数 I とする．また，試聴可能な部分は曲の代表的な部分であると考え，この音響信号から楽曲の音響的特徴を表現するための音響ベクトルを抽出する．

3.2 評価方法

2,705 曲の音響信号とレビューとのペアを 5 つのグループにわけ，5 つを学習と評価の両方に利用する closed テストと，4 つを学習データ，1 つを評価データとして 5-fold クロスバリデーションを行う open テストを行った．学習データから推定された変換行列 W に，評価データである曲 m の文書ベクトル $\mathbf{t}_m$ をかけて推定される音響ベクトル $W\mathbf{t}_m$ と真の音響ベクトル $\mathbf{a}_m$ との 2 乗誤差 $e_{m,m} = ||\mathbf{a}_m - W\mathbf{t}_m||^2$ が ε 以内であれば正解とする．また，別の曲 l の音響ベクトル $\mathbf{a}_l$ と $W\mathbf{t}_m$ との 2 乗誤差 $e_{l,m} = ||\mathbf{a}_l - W\mathbf{t}_m||^2$ も利用して，変換行列 W の推定性能を評価するために，情報検索システムの評価に利用される再現率，適合率の考え方を取り入れる．再現率は，評価データの曲数に対して，正解と出力された曲数の割合であり，(7) 式のように定義する．

$$\text{再現率 } (R) = \frac{e_{m,m} \leq \varepsilon\text{をみたす曲数}}{\text{評価データの曲数}} \tag{7}$$

一方，適合率は，評価データを入力したとき，2 乗誤差が ε 以内であった曲数に対して，どれだけ正解が含まれているかという正確性の指標として，(8) 式のように定義する．

$$\text{適合率 } (P) = \frac{e_{m,m} \leq \varepsilon\text{をみたす曲数}}{e_{l,m} \leq \varepsilon\text{をみたす曲数}} \tag{8}$$

最終的に，この二つを統合した F 値

$$\text{F 値} = \frac{(\beta^2+1)RP}{\beta^2 P + R} \tag{9}$$

を用いる ($\beta = 1$)．ε を変化させ，楽曲の音響ベクトルと文書ベクトルがどれだけ正確に，また網羅的に変換行列 W によって関連付けられているかについて検証する．

3.3 実験結果

変換行列 W のサイズを 1,024×1,024 に固定し，ε を変化させたときの再現率と適合率を図 2 に示す．ε を大きくすると適合率は下降し，再現率は上昇する．このとき F 値の最大値は，closed データによる評価で $\varepsilon = 3.1 \times 10^{-6}$ のときに 0.628，open データによる評価で $\varepsilon = 3.6 \times 10^{-6}$ のときに 9.96×10^{-3} であり，open データでは低い F 値を確認した．

図 3 は変換行列のサイズを変化させたときの F 値の最大値を示す．行列サイズを大きくすることによって F 値は上昇した．また，closed データによる評価と比べて，open データによる評価で関連付け性能が低いことを確認した．この open データに適応できない原因の 1 つとして，変換行列 W の学習が十分でないことが考えられる．使用した 2,705 曲にさらに曲を追加して学習データ量と関連付け性能との関係について調査する必要がある．今回は楽曲に対して単

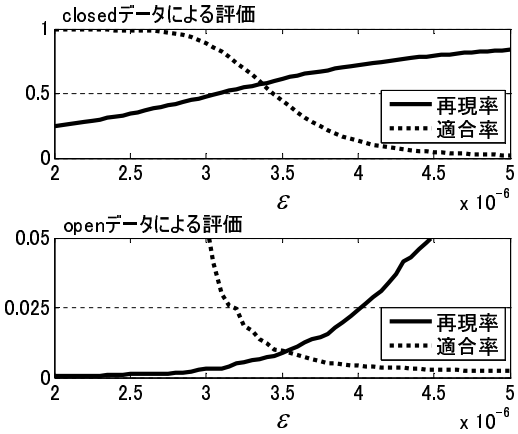

図 2: ε による再現率，適合率の変化

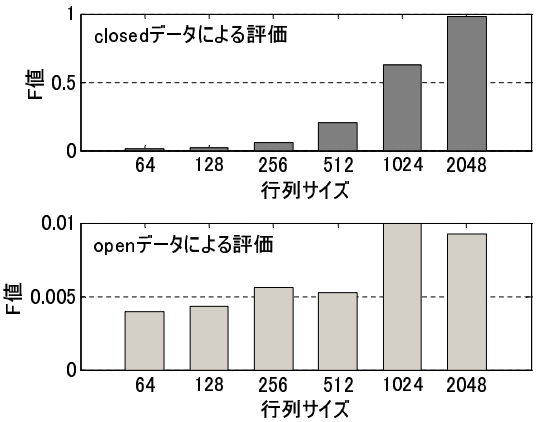

図 3: 行列サイズによる F 値の変化

一のレビューを使用したが，楽曲に対する複数のレビューを大量に集めること，歌詞等のテキストデータを加えることにより，文書ベクトル抽出のための学習データを増やす必要性も考えられる．また，提案した音響ベクトルによって楽曲の音響的特徴をとらえることが十分であるかについても検討する必要がある．

行列サイズ 2,048×2,048 の closed データによる評価で F 値 0.981 が得られた．すなわち，文書ベクトル，音響ベクトルの次元が 2,048 のときに 2,705 曲を 98.1%の精度で関連付けることが可能である．2,048 次元の文書ベクトルはどのような形態素の寄与によって構成されているか調査することも必要である．

4 まとめと今後の展開

楽曲のレビューに出現する形態素を TF-IDF によって重みづけした文書ベクトルと，楽曲の音響的特徴を表現するためにクロマベクトルの頻度分布を利用した音響ベクトルを提案し，線形変換で関連付けることを試みた．2,705 曲のレビューと音響信号を使用して関連付けの性能について評価したところ，closed データによる評価で F 値は最大 0.981 が得られたが，open データによる評価で関連付け性能の低さが確認された．今後は，関連付けの学習に必要な曲数の検討，また文書ベクトルと音響ベクトルの表現方法に関して再検討する予定である．

参考文献

[1] 池添ら，“ 音楽感性空間を用いた感性語による音楽データベース検索システム ”，情処学論，vol.42, no.12, pp.3201-3212, 2001.

[2] 竹村ら，“ 言語と心理の統計 ”，岩波書店，pp.139-143，2003.

[3] 音楽ダウンロード・メガサイト Mora, http://mora.jp/

[4] 後藤真孝，“ リアルタイム音楽情景記述システム：サビ区間検出法 ”，情報処理学会研究報告，2002-MUS-47-6, Vol.2002, No.100, pp.27-34, 2002.