アーティスト評判分析に有効な特徴表現の抽出

山田 達史　　松本 和幸　　吉田 稔　　北 研二
徳島大学工学部　知能情報工学科

1．はじめに

情報社会といわれる近年では，誰しもが情報資源を頼りにする生活をしている．たとえば，気になる商品があればインターネットで調べ，口コミを見て，インターネットショッピングサイトでそのまま購入し，その商品のレビューの書き込みをしたり，日記に綴るかのように近況をFacebookやTwitterなどにアップロードし，遠隔コミュニケーションをとることもできる．本研究では大容量のテキストデータとして，Twitterからtweet（つぶやき）を収集する．tweetは一つの記事に対する文字数が制限（140字以内）されており，従来のブログ（Weblog）よりも更新回数が多く，利用者も多いため，多種多様なテキストデータであるという特徴が挙げられる．Twitterなどのマイクロブログにおいて，テキストデータは既存の形態素解析で解析するのは困難であるとされている．たとえば，“きゃりーぱみゅぱみゅのもったいないとらんど聞いてるなう”を形態素解析したものを示す．

［きゃ/動詞，り/助動詞，―/名詞，ぱみゅぱみゅのもったいないとらんど/名詞，聞い/動詞，てる/動詞，な/助詞，う/感動詞］

［曲名］や［アーティスト名］が既存辞書には存在しないため正しく分割できていない事は明らかである．また，“なう”といったtweet特有の表現も解析できない．

また，従来の評判分析ではアーティストに関しては“positive”，“negative”に分類し，評価を行うが，アーティストに関して，goo評判分析を行うと表1に示すとおりになった．positiveに示す値は，positiveでの割合である．

表1　goo評判分析による評価判定（2014年）

アーティスト名	positive	対象となった語彙
氷川きよし	*0.86*	大好きな，好きな
サザンオールスターズ	*0.84*	大好きな，待ち望んだ
少女時代	*0.85*	良い，可愛い
きゃりーぱみゅぱみゅ	*0.87*	笑った，いい
ももいろクローバーZ	*0.81*	初めて見ました，自身あり

表1から，positiveの評価の差が非常に少なく，対象となった語彙にも差がみられないため，従来手法でアーティストごとの差を見出すことはできない．

各アーティストにおいて特徴的な表現を抽出することができれば，アーティスト評判分析に有効な情報が得られると考えられる．たとえば，関連の深い［場所］を抽出できることで，コンサートでの情報を収集できたり，関連の深い［曲名］を抽出することで，“可愛い○○”や，“面白い○○”といった，評価表現が含まれた曲名での誤差を減らすことができる．本研究では，n-gram出現頻度に基づき，特徴表現を抽出する．さらに，特徴表現に後続する単語を素性として，機械学習をおこなうことで各アーティストに関連する特徴表現を分類する．

2．関連研究

Twitterを使った研究として，井戸木 [1]は楽曲情報の抽出する研究をおこなっている．Twitterでは自然言語処理で多用される形態素解析による分析が困難であるため，曲名とアーティスト間の関係に着目

し，パターンマッチングによる抽出をおこなうというものである．フォーマット例を表 1 に示す．また，水岡ら [2]はマイクロブログを用いて感情表現収集をおこなっている．水岡らの研究では感情を「かっこいい」「かわいい」「泣ける」「笑える」の 4 つ分類した研究である．代表的な感情表現と，新たに感情表現となりうる表現を比較し，収集している．

3． 提案手法

本研究では，以下の手順により，アーティストに関連する特徴表現を抽出する．

STEP1．収集及びフィルタリング

STEP2．n-gram 出現頻度による特徴表現抽出

STEP3．特徴表現に後続する素性を抽出

STEP4．SVM による学習及び分類

表 2　手作業による分類

分類	特徴表現
イベント	FNS 歌謡祭，紅白，夏のバカ騒ぎ
人／グループ	れにちゃん，あーりん，しおりん，モノノフ
場所	日産スタジアム，西武ドーム
曲名	GOUNN

3.1 教師データの作成

3.1.1.　収集及びフィルタリング

Twitter API[3]を使用して，“ももいろクローバーZ”に関する tweet を 2013 年 11 月～12 月の 1 か月間収集した. tweetの中には, 宣伝を含むものや全く同じtweetが含まれている．本研究では “ヤフオク”, “入札”, “オークション”, “送料無料”, “デリヘル”, “フォロー”, “合コン”, “拡散希望” を含む tweet を除去する．また，短縮 URL，リプライ時に挿入される相手側の ID の削除を行った．

3.1.2.　n-gram 出現頻度による特徴表現抽出

フィルタリングした tweet の文字 n-gram の頻度統計をとる．4-gram から 15-gram までを抽出し，頻度上位のものを特徴表現とする．つぎに，特徴表現のうち，表 2 に示すような分類に属する表現について，人手によりラベル付けをおこなった．たとえば，“モノノフ”というのは，“ももいろクローバーZ”のファンの総称であるが，［人／グループ］のラベルを付与した．さらに，特徴表現を tweet 内から検索し，それに後続する内容語を組成として抽出した．

3.2 SVM による分類

前節で述べた手順の通りに準備したデータを，多値分類が行える SVM multiclass[5]を用いて学習させ, 分類モデルを作成する．

4． 実験と考察

提案手法の評価をおこなうため，収集した tweet に対して，提案手法により特徴表現について 4 種類の分類をおこなった．対象となるアーティストは，”ももいろクローバーZ“である．10 分割交差検定法を用いて分類結果の評価をおこなう．各分類ごとの tweet 数とその正解率を，表 3 に示す．

表 3　SVM による実験結果

分類	tweet 数	正解率[%]
イベント	2191	86.52
人／グループ	13107	85.74
場所	11311	81.90
曲名	2145	58.00

以上の結果から，[曲名]の正解率が比較的低くなった．この理由として，後続する素性表現にばらつきがあったためだと思われる．たとえば，“リズム，流れる”の場合，有効な素性となるが，“好き，良い”などの表現（評価表現）が多く出てしまっていた．こうした表現は，分類に有効ではない素性だと考えられる．

5． 今後の課題

今後の課題として，他のアーティストについても実験を行う必要がある．さらに，特定できる素性ではないものは除去する必要がある．

参考文献

1. 井戸木良. Twitter における楽曲情報の抽出. 2013

2. 水岡良彰，鈴木優．マイクロブログを用いた感情表現収集.

3. Twitter Developers.　https://dev.twitter.com/.

4. Twitter.　https://twitter.com/.

5. SVM multiclass
http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html