女性グループの歌詞の計量テキスト分析

小林佳織
東洋大学社会学部

狩野恵里奈
東洋大学社会学部

鈴木崇史
東洋大学社会学部

女性グループの曲は,現在の日本全体の楽曲売り上げにおいて大きな部分を占め,2011年度のオリコン年間トップ50においては17曲がランクインする結果となっている．また，流行歌の歌詞を理解することはその時代のポピュラーカルチャーを社会学的に理解するために重要な意義をもつ．そこで本研究では，過去35年間のオリコン年鑑のデータをもとに女性グループ34組439曲を対象とし研究を行った．歌詞のテキストファイルを作成し，形態素の頻度を計量，主成分分析とランダムフォレスト機械学習法を適用した．その結果，ヒット歌手，各年代にそれぞれ違いがあることを明らかにした．本研究は女性グループを対象とし，歌手や年代の違いを知るための歌詞の重要性を考える．

1. はじめに

いつの時代にも人気のグループ歌手は存在し,そのグループの活動形態はアイドルやバンド,声優によるアニメキャラクターに扮したものなど多岐にわたる．時代ごとに流行歌をみていくと，そのなかには女性グループのものが数多く含まれている．古くはキャンディーズやピンク・レディなどが人気を得た．その後 SPEED,モーニング娘。，AKB48 がヒット曲を生み出した．近年では AKB48 のヒット以降女性グループに注目が集まり，多くの女性グループが登場し,「アイドル戦国時代」と言われている[1].

流行歌は時代や社会の表象として重要な研究対象となってきた.特に女性グループの曲は,現在の日本全体の楽曲売り上げにおいて大きな部分を占める．2011年度のオリコン年間トップ50においては17曲がランクインする結果となり，またアーティストトータルセールストップ5でも3組が入った[1]．よってこれを分析することは，新たなポピュラーカルチャーを理解するために重要な意義をもつ．とりわけ歌詞は，楽曲を構成する上で重要な役割を果たしており,その時々の時代を社会学的に理解するために有用である[2].

一方，近年，言語処理技術の発達とデータベースの蓄積によりテキストマイニングの射程はますます広がっている．従来歌詞の研究は見崎[3]の質的研究や伊藤[4]の特定の歌手に注目したものがあったが，近年では言語処理技術の発達により，テキストマイニングを利用した研究も行われている[5].

このような背景のもと本研究では 1977 年から 2011 年の過去 35 年間にわたる女性グループによる楽曲(34組439曲)をもとに分析を行う．歌詞のテキストファイルを作成し，形態素の頻度を計量，主成分分析とランダムフォレスト機械学習法を適用し探索的に分析を行う．その結果をもとに女性グループの歌詞による，歌手や年代の違いの有無を明らかにする.

2. データと分析手法

2. 1 データ

本研究の調査対象は1977年から2011年の間にオリコンの年間売り上げトップ 50 位[6;7]にランクインした2人以上の女性だけで構成された日本のグループとした．各グループの 1977 年 1 月 1 日から 2011 年 12 月 31 日までに発売されたすべての曲[2]を対象に分析を行なう[3]．分析対象は34組439曲[4]となった.

[1] www.oricon.co.jp/music/special/2011/musicrank1219/index09.html#topphoto (最終アクセス2013年1月9日)

[2] 歌詞が入手できず,対象外となった曲が12曲あった．また配信限定曲は除いた.

[3] 1枚しかシングル曲が無い場合は分析の対象外とし，ランクインした時点で2人以上のグループである場合には対象とした.

[4] 両A面の曲は2曲とも対象とした.

2. 2 分析手法

形態素の出現頻度の計量

歌詞のテキストファイルは，2011 年 10 月から 2012 年 11 月にかけて「歌ネット[5]」,「うたまっぷ[6]」,「歌 GET!![7]」を参照し，作成した．タイトル, スペース, 歌手, 作詞, 作曲, 記号「※」と「() を含む（）内文字」,「,」を削除した．形態素解析には MeCab[8]を用いた．全形態素の出現頻度を計量し，全語彙，内容語（名詞・動詞・形容詞・副詞・連体詞・感動詞・フィラー）[9]，機能語（接続詞・助詞・助動詞・記号）それぞれ，上位 20 形態素の頻度を観察した．

主成分分析

テキストを行，各形態素の相対頻度を列とするテキスト - 特徴量行列を作成，分散共分散行列を作成し，主成分分析を適用した[8]．これによって，主成分を抽出し，テキストに影響を与える要因を探索的に分析した．同時に，テキストの位置関係を可視化することで，グループのまとまりを観察した．

ランダムフォレストによる分類実験

機械学習による分類実験にランダムフォレスト[9]を用いた. テキスト - 特徴量行列から 1000 個のブートストラップを作成し，列の平方根をランダムサンプリングで抽出した．ブートストラップデータの 2/3 を学習に, 1/3 をテスト用データに利用した．

分類クラスは歌手ごと（34 クラス），20 曲以上の曲数がある歌手ごと（7 クラス) ，年代ごと（4 クラス）とし，以下の 3 つの実験を行った．

実験 1：全歌手ごとの実験

実験 2：20 曲以上の歌手の歌手ごとの実験[10]

実験 3：年代ごとの実験[11]

分類結果を観察することで，データセットに対して，特別な特徴を持つ（分類性能の高い） 歌手，特徴の少ない（分類性能低い）歌手を明らかにすることができる[2]. 評価実験には, 精度, 再現率，*F1* 値を用いた[10].

3. 結果と考察

3.1 形態素の出現頻度の計量

表 1 はそれぞれの歌手について分析対象曲数, 人数, オリコン年間トップ 50 位にランクインした曲数，1 曲あたりの延べ語数を示したものである．楽曲数は 2 曲から 47 曲の範囲にあり，1 曲あたりの延べ語数は 122 語から 691 語の範囲にある. 全テキストにおける延べ語数は 106,487, 異なり語数は 10,001 であった. わらべ, WINK, PUFFY，NMB48 は相対的に曲が短いのに対し, BENNIE K，らきすた[12]，放課後ティータイムは相対的に曲が長いのが特徴である．BENNIE K に関してはほかのグループに比べ全体的に英語の歌詞が多くなっていることが影響していると推察する．また放課後ティータイム，らきすたは記号が曲中に多く含まれ，歌詞を強調するために一語ずつになっているからと推察する．GO-BANG'S，Gorie with Jasmine & Joann，モーニング娘。, プッチモニ, 放課後ティータイムは標準偏差，変動係数ともに高い結果となった．

表 2 はすべての曲を対象に形態素の出現頻度を全語彙, 内容語, 機能語ごとに上位 20 位までを計量した結果である．全語彙でみると「あなた」以外はすべて機能語となった．内容語では一人称が「私」と英語の「I」，二人称が「あなた」と「君」と人称が混在していた. また,「恋」,「愛」といった恋愛に関する言葉が上位に来ていることからラブソングが多いと推察する．

---

[5] www.uta-net.com/

[6] www.utamap.com/

[7] www2.kget.jp/

[8] mecab.sourceforge.net

[9] 内容語，機能語ごとに分析を行う際，MeCab では英単語はすべて名詞に分類されてしまうため，全編英語の歌詞の曲 PUFFY の 2 曲とピンク・レディの 1 曲を分析の対象外とした．

[10] AKB48，PUFFY，SPEED，WINK，モーニング娘。，ピンク・レディ，プリンセスプリンセスの 7 組．曲数はそれぞれ最新のもと古いものから 10 曲ずつ選び計 20 曲とした．

[11] 曲数を一番少ない 1970 年代に合わせ各年代 34 曲とし，曲はランダムに抽出した．

[12] CD 発売時の名義は泉こなた (平野綾)，柊かがみ (加藤英美里)，柊つかさ (福原香織)，高良みゆき (遠藤綾) であったが便宜上，この声優たちが演じたアニメのタイトルを使用する．

表 1　基礎データ

	曲数	人数	50位以内	延べ語数		
				平均	標準偏差	変動係数
AKB48	23	16	13	281.87	65.55	22.15
あみん	7	2	1	191.57	30.36	15.85
BENNIE K	14	2	1	439.29	60.83	13.85
キャンディーズ	18	3	6	185.28	38.31	20.68
フレンチ・キス	4	3	1	222.50	46.77	21.02
FUNK THE PEANUTS	5	2	1	319.00	28.43	8.91
GO-BANG'S	10	2	1	258.30	104.79	40.57
Gorie with Jasmine & Joann	3	3	2	251.67	87.29	34.69
Kiroro	18	2	2	197.83	42.46	21.46
KIX-S	14	2	1	204.07	49.66	24.34
Mi-Ke	11	3	1	186.45	38.80	20.81
ミニモニ。	18	4	5	290.61	84.02	28.91
モーニング娘。	47	9	9	278.30	98.52	35.40
NMB48	2	16	2	175.50	0.71	0.40
Not yet	3	4	3	262.67	67.11	25.55
おニャン子クラブ	10	18	3	210.60	36.96	17.55
Perfume	17	3	1	219.76	74.08	33.71
PINK SAPPHIRE	7	4	1	216.00	19.05	8.82
ピンク・レディ	24	2	12	194.75	47.35	24.32
プリンセスプリンセス	22	5	5	206.68	37.92	18.34
PUFFY	39	2	4	176.64	54.60	30.91
プッチモニ	5	3	2	327.00	166.59	50.95
らきすた	3	4	1	547.67	50.90	9.29
SKE48	7	16	4	297.00	89.91	30.27
SPEED	21	4	12	313.76	84.57	26.95
Sugar	4	3	1	221.75	75.16	33.89
タンポポ	8	4	1	205.38	26.42	12.87
放課後ティータイム	12	5	5	325.75	112.56	34.55
うしろゆびさされ組	6	2	3	216.67	24.74	11.42
わらべ	3	3	2	139.33	15.82	11.36
Whiteberry	11	5	1	241.73	58.37	24.15
WINK	25	2	6	184.80	50.13	27.13
やまだかつてないWINK	2	2	1	212.00	7.07	3.34
ZONE	16	4	1	260.38	73.87	28.37

131415

表 2　出現形態素上位 20 位

	全語彙		内容語		機能語	
	語	回数	語	回数	語	回数
1	の	4111	し	770	の	3726
2	て	3764	あなた	695	て	3498
3	に	3079	てる	503	に	3079
4	を	2063	私	468	を	2063
5	は	2044	人	426	は	2044
6	た	2006	の	385	た	2006
7	が	1868	い	384	が	1868
8	！	1685	君	381	！	1685
9	ない	1603	この	375	で	1507
10	で	1510	恋	328	も	1459
11	も	1459	いる	315	ない	1378
12	よ	1034	愛	300	よ	1026
13	な	1014	夢	277	な	1014
14	だ	818	いい	275	だ	818
15	し	816	て	266	と	753
16	と	762	こと	264	う	669
17	あなた	695	I	261	？	591
18	う	481	今	260	ね	550
19	？	591	ん	255	…	545
20	ね	550	れ	244	か	531

3.2 主成分分析

図 1 はテキスト特徴量行列から作成した分散共分散行列に主成分分析を適用した結果である．第一主成分，第二主成分ともに高い絶対値をもっているのが SKE48，プッチモニ，モーニング娘。，AKB48，SPEED などである．その中でも SKE48 とプッチモニはほかのものと比べ外れている．図 2 は第一主成分に寄与する上位 20 語である．名詞，動詞，助動詞などが混在しており，軸の解釈自体は自明ではない．

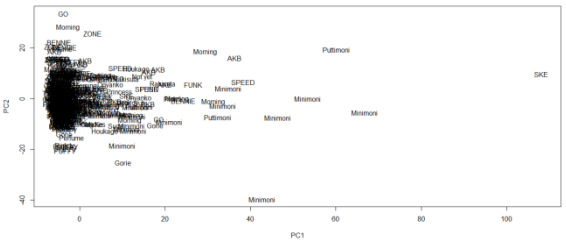

図 1　主成分分析の結果

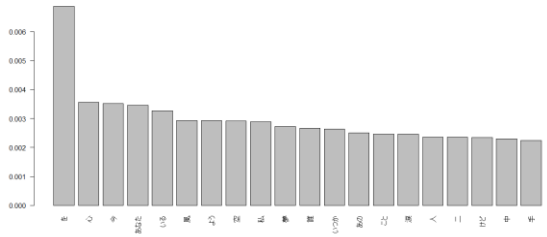

図 2　第一主成分に寄与する上位 20 語

3.3　ランダムフォレストによる分類実験

表 3 は 3 つの分類実験の結果である．実験 1 は低い結果となり，再現率，*F1* 値は数値を得ることができなかった．実験 2，3 では相対的に良い結果を得た．

表 3　分類実験の結果

	精度	再現率	F1 値
実験 1	20.120	−	−
実験 2	57.821	61.763	58.027
実験 3	52.089	52.201	50.426

実験 1 の歌手ごと（34 クラス）の分類では AKB48, BENNIE K, PUFFY, Perfume, SPEED, WINK, モーニング娘。がエラー率が低い結果となった．この結果は，これらの歌手が特別な特徴をもつ歌手と示唆される．ただし，曲数の影響も大きいと示唆される．

表 4 は 20 曲以上曲のある歌手の曲数を 20 曲に限定し，歌手ごと（7 クラス）に分類した結

[13] PINK SAPPHIRE，GO-BANG'S ，わらべはそれぞれ 1 曲，ピンク・レディは 2 曲，Sugar は 7 曲の歌詞が入手できず含まれない曲数となっている．

[14] FUNK THE PEANUTS rated R，PUFFY×東京スカパラダイスオーケストラ，桜高校軽音部，ミニハムず，バカ殿様とミニモニ姫。，ミニモニ。と高橋愛＋4 KIDS 名義を含む．

[15] 人数は解散時，または最新曲時のものとした．

果である．実験 1 ではエラー率の低かったモーニング娘。が，曲数を調整するとエラー率が一番高い結果となった．分類もすべての歌手に分類されていることから色々な歌手の要素を含んでいるが, 独自の特徴は少ない歌手と推察する．そのほかの歌手についてはエラー率に大きな違いはなかった．

表 4　全特徴量を用いた 20 曲以上の歌手の結果

	AKB48	PUFFY	SPEED	WINK	モーニング娘。	ピンク・レディ	プリンセスプリンセス	エラー率
AKB48	13	2	0	0	1	3	1	0.350
PUFFY	1	11	0	4	0	3	1	0.450
SPEED	1	1	14	0	1	1	2	0.300
WINK	0	3	0	14	2	0	1	0.300
モーニング娘。	2	3	2	2	6	4	1	0.700
ピンク・レディ	0	3	0	2	0	14	1	0.300
プリンセスプリンセス	0	1	1	5	0	2	11	0.450

表 5 は年代ごと（4 クラス）の分類結果である．1990 年代のエラー率 0.794 以外は，エラー率は低い結果となった．このことから各年代とも一定の特徴があるといえる. 特に 1970 年代は 0.088 という低い結果であり, はっきりした特徴があるといえる． ただし，1970 年代はグループが 2 組しかなくほかの年代と比べ少ないため, 特徴が明確にでていることが示唆される．

表 5　年代ごとの結果

	1970年	1980年	1990年	2000年	エラー率
1970年	31	2	0	1	0.088
1980年	7	15	9	3	0.559
1990年	9	12	7	6	0.794
2000年	5	5	3	21	0.382

5. おわりに

本研究では過去 35 年にわたる女性グループの曲を探索的に分析した．形態素の出現頻度の計量では恋愛に関する形態素が上位に来ることを明らかにし，ラブソングが多いこと示した．

ランダムフォレスト機械学習法による分類実験では 20 曲数以上ある歌手に限定した実験において，それぞれの歌手に違いがあることを明らかとした．また各年代にも違いがあり，時代ごとに異なることを示した．よって歌詞は歌手や年代の違いを明らかにする重要な要素である．

年代ごとの分析では 1970 年代が少ないなどデータの偏りがあり，時代ごとの特徴を読み取るには十分なものとならなかった．また女性グループに限定したことで歌詞全体の研究とはなっていない. 今後は男性歌手やソロ歌手も含め, 他の特徴量も検討し研究を行いたい.

謝辞

本研究は，科研費（若手（B），2011-2014 年）「計算文体論による多種メディアテキスト解析（研究代表者:鈴木崇史）」(課題番号:23700288) の助成を受けたものです．ここに記して，謝意を表します．

文献

[1] 岡島紳士・岡田康宏．グループアイドル進化論: 「アイドル戦国時代」がやってきた！, 毎日コミュニケーションズ，東京，2011.

[2] 三田宗介．近代日本の心情と歴史：流行歌の社会心理史，講談社，東京，1978.

[3] 見崎鉄． J ポップの日本語: 歌詞論，彩流社，東京，2002.

[4] 伊藤雅光．ユーミンの言語学(40): ユーミン語彙の品詞比率から文体をさぐる，日本語学, 20(1)，74-82，2001.

[5] 細谷舞・鈴木崇史．女性シンガーソングライターの歌詞の探索的分析, じんもんこん 2010: 人文科学とコンピュータシンポジウム論文集, 195-202，2010.

[6] コンフィデンス年鑑（1970-1979）．オリコン・エンタテインメント，1970-1979.

[7] オリコン年鑑（1980-2011）．オリコン・エンタテインメント，1980-2011.

[8] 金明哲． R によるデータサイエンス: データ解析の基礎から最新手法まで，森北出版，東京，2007.

[9] Leo, Breiman, Random forests, *Machine Learning*, 45(6), pp. 5-23, 2001.

[10] 徳永健伸．情報検索と言語処理，東京大学出版会，東京，1999.