言語解析ソフトウェアによる流行歌の歌詞とそこに反映される世相の分析

平井　健斗（静岡大学大学院総合科学技術研究科情報学専攻）

白井　靖人（静岡大学情報学部行動情報学科）

本稿では，機械学習の手法を用いて，1945 年から 2019 年までの流行歌 663 曲を対象に、歌詞からその歌が発表された時期の推定を行った．歌詞の分析を行うことで、楽曲の発表年の推定ができうること，ジャンルの推定については，（今回調査した範囲では）その効果があまり見られなかったこと，景気拡張期間を広げるような解釈が好まれるということが確認できた．また，歌詞の中に最も多く登場した名詞“夢”の使われる文脈に注目し，歌の発表年による変化の有無を調査したところ、ネガティブな意味合いでの使用頻度が減り、ポジティブな意味合いでの使用頻度が増えたことがわかった．

Analysis of popular song lyrics and the social situation reflected in them by language analysis software

Kento Hirai (Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University)

Yasuto Shirai (Department of Behavioral Information, Faculty of Informatics, Shizuoka University)

In this paper, we applied a machine learning technique to estimate the release time of Japanese popular songs from their lyrics. We tested it on 663 songs released between 1945 and 2019. By analyzing the lyrics, it is possible to estimate the release year of the song, and regarding the estimation of the genre, the effect was not so much seen (within the range investigated this time), and the interpretation that extends the economic expansion period is preferred. In addition, focusing on the context in which the most frequently used noun "夢 (dream)" is used, we investigated whether there was a change depending on the year of release. It was found that the number of uses in the negative sense decreased and the number of uses in the positive sense increased.

１．はじめに

「歌は世につれ．世は歌につれ．」と言われるように，流行歌は世相を反映するものとされている．そこで，言語解析ソフトウェアを用いて流行歌の歌詞の特徴の変化を明らかにすることによって，そこに反映される世相の分析を試みた．

世の中は変化を繰り返すものであり，十年一昔という表現があるように，何年か経つとその姿や人の考えもが大きく変化する．このような変化については，経済学の分野でも，長期，中期，短期の循環の存在が指摘されている．こちらの循環は，一連の事象の発生が繰り返し発生するというもので，コンドラチェフサイクル，ジュグラーサイクルなどが知られている．[1]

本稿では，歌詞からその歌が発表された時期を推定することを試みる．発表時期としては，いつ発表されたかを推定する発表年による分類，発表されたときの世の中の時の様子として推定する景気動向による分類を取り上げた．景気動向は拡張と後退を繰り返すものであり，一連の事象の繰り返しに対応するものとして採用した．

発表年については，調査対象期間を一定の長さ（1～20 年）で区切り，そのうちのどこに属するかを推定し，区切りの長さと正答率の関係を調査する．その際曲のジャンルを考慮することの効果についても検証する．景気動向については，歌の発表された月が拡張期間と後退期間のいずれに属するかを推定し，その正答率を評価する．拡張期間と後退期間の別は，政府が発表する景気基準日付をもとに判断する．

発表時期の推定に加え，特定の語（ここでは，“夢”）の使われる文脈に注目し，歌の発表年による変化の有無を調査した．

歌詞データを分析する研究にとしては，[2]，[3]など，特定のアーティストやグループ形態に着目する研究が存在する．また，長期間にわたる調査としては，35 年間（1978～2012 年）の日本レコード大賞および優秀作品賞の受賞作品を対象に，当時の社会背景との関係性を分析した[4]がある．本研究では，特定のアーティストやグループ形態

にはこだわらず、2019 年までの 70 年分のヒット曲を対象に，分析を行った．

2 節で発表時期推定のための研究方法について説明し、3 節で発表年（ジャンルを考慮しない場合，ジャンルを考慮する場合）と発表時の景気動向に関する推定の結果について述べる．4 節では、歌詞中に最も多く登場する単語「夢」の文脈を調査する．5 節では，まとめと今後の課題に触れる．

なお，機械学習分野の用語に倣い，以下において，発表時期を推定することについて“カテゴリに分類する”という表現を併用する．

２．研究方法

発表時期の推定にあたって，本稿で用いた研究方法について以下に述べる．言語解析には，言語解析ソフトウェア KH coder[5]を用いた．

1. 1945～2020 年の各年のシングルヒット年間上位 10 曲を，web サイト「年間シングルヒット曲」[6]から抽出し，各歌に関する情報とともに，2.以降の調査対象とする．
2. 1.で抽出した歌の歌詞を web サイト「歌マップ」[7]から取得する．
3. 1.で取得した各歌の情報をもとに、該当するカテゴリ（分類の際の正解データ）を手作業で追加する．
4. 3.の結果を KH Coder に読み込み，「ベイズ学習による分類」の機能を使って発表時期を推定する．
5. 「交差妥当化」によって，調査対象となる全ての歌について分類を行い，正答率を求める．

調査対象の抽出において，情報源である[6]の成約から，1945～1949 年についてはほかと異なる扱いをしている．この期間については，5 年間の上位 10 曲をえらび，それを 1950 年に発表されたものとして扱っている。

また，2.において，歌詞の情報が入手できない，あるいは，歌詞が全編英語であるなどの理由によって，調査対象から除外したケースがある．調査対象の歌が 10 曲に満たない年は，表 1 のとおりである．この結果，663 曲を今回の調査対象とした．

表 1 取得できた歌詞データ数が 10 曲未満である年

曲数	年
6曲	1952 1953
7曲	1958 1961 1964 1968
8曲	1954 1962 1963 1965 1983 1999
9曲	1950 1951 1955 1956 1957 1960 1969
	1970 1976 1978 1987 1989 1990 2004

4.で用いるベイズ学習による分類では，ナイーブベイズモデルが用いられている．歌詞に登場する語とその使用頻度を学習させ，それをもとに未知の歌の発表年を類推させる．一連の語 W を含む文書（ここでは，歌詞）が，カテゴリ C に属する確率 p(C|W)は，ベイズの定理より以下のようにあらわすことができる．

$$p(C|W) = p(W|C)p(C) / p(W)$$

すべてのカテゴリについてこの値を計算し，その値が最大となるカテゴリに文書を分類する．

３．推定

3-1.発表年の推定

今回の対象期間である 1950～2019 年の 70 年間を一定間隔に区切り，区切られた期間のそれぞれをカテゴリと見立てて，歌の分類を行う．実際にその歌が属するカテゴリに分類されれば正答であり，対象中で正しく分類された歌の割合を正答率という．

区切りの幅を 5，10，15，20 年と 4 通りに変化させた場合の正答率と，分類結果の内訳を表 2 に示す．区切リの幅が広いほど正答率は高くなるのは当然の結果であり，少なくともこの点においてこの方法は妥当であることが確認できる．

カテゴリをランダム選んで回答したとすると，正答率は 5 年区切りで 7.1%（=1/14），10 年で 14.3%，15 年で 20.0%，20 年で 25.0%となる．この方法による結果はランダムな回答の場合よりも高い正答率を示しており，歌詞を分析することは発表年の推定において有効であることが確認できる．

また，正しく分類できなかったデータの数に着目すると，実際の年代よりも前の年代に分類されたデータの数よりも，後ろの年代に分類されたデータの数が多いことがわかる．このことは，ヒット曲の歌詞は時代の流れよりも先行する傾向を持つと解釈される．

表 2 ベイズ学習による発表年推定の結果

区切り方	カテゴリ数	正答率	※1	※2	※3
5年区切り	14	25.9%	132	163	368
10年区切り	7	48.7%	92	323	248
15年区切り	5	56.9%	69	377	217
20年区切り	4	65.5%	52	434	177

※1：実際のデータよりも前の年代に分類されたデータの数

※2：実際のデータと同じ年代に分類されたデータの数（正答）

※3：実際のデータよりも後ろの年代に分類されたデータの数

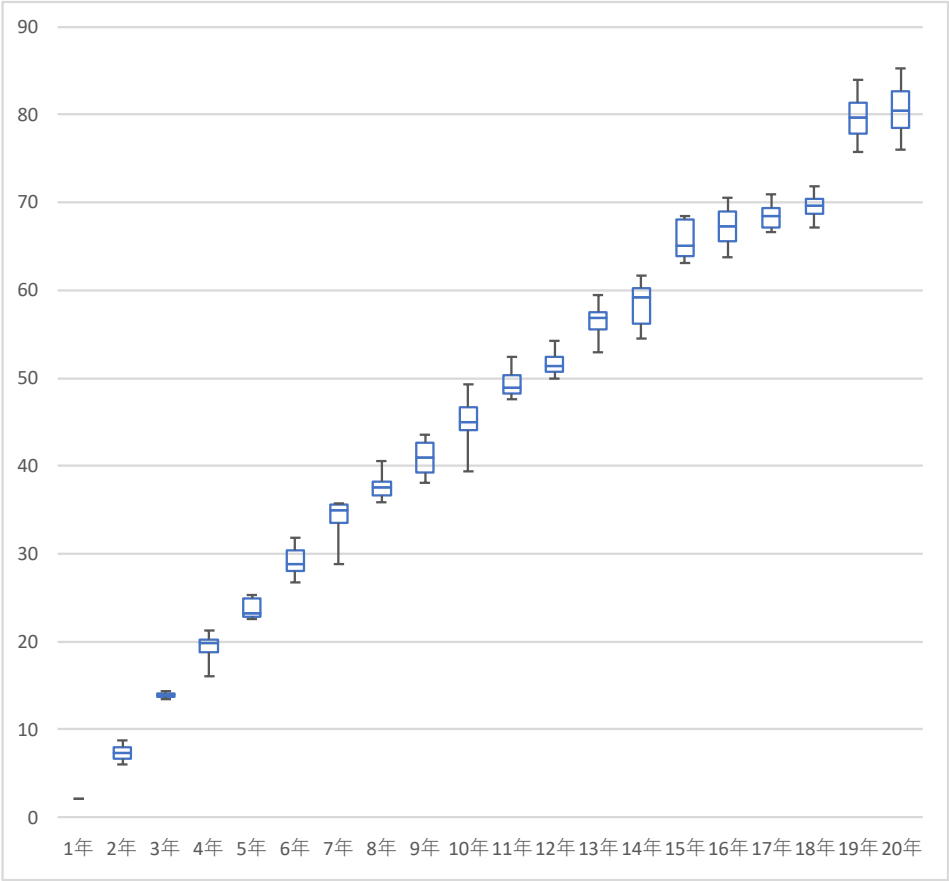

図 1　区切り年数と正答率の関係

表 4　景気基準日付(一部)

景気基準日付

循環	谷	山	谷	期間			(参考)四半期基準日付	
				拡張	後退	全循環	山	谷
第1循環		1951年6月 (昭和26年6月)	1951年10月 (昭和26年10月)		4か月		1951年4-6月 (昭和26年4-6月)	1951年10-12月 (昭和26年10-12月)
第2循環	1951年10月 (昭和26年10月)	1954年1月 (昭和29年1月)	1954年11月 (昭和29年11月)	27か月	10か月	37か月	1954年1-3月 (昭和29年1-3月)	1954年10-12月 (昭和29年10-12月)
第3循環	1954年11月 (昭和29年11月)	1957年6月 (昭和32年6月)	1958年6月 (昭和33年6月)	31か月	12か月	43か月	1957年4-6月 (昭和32年4-6月)	1958年4-6月 (昭和33年4-6月)
第4循環	1958年6月 (昭和33年6月)	1961年12月 (昭和36年12月)	1962年10月 (昭和37年10月)	42か月	10か月	52か月	1961年10-12月 (昭和36年10-12月)	1962年10-12月 (昭和37年10-12月)
第5循環	1962年10月 (昭和37年10月)	1964年10月 (昭和39年10月)	1965年10月 (昭和40年10月)	24か月	12か月	36か月	1964年10-12月 (昭和39年10-12月)	1965年10-12月 (昭和40年10-12月)

(以下省略)

区切の幅と正答率の関係を示したグラフを図1に示す。

全体として区切りの幅が大きいほど正答率が上がることが確認できるが，14年から15年，18年から19年にかけて正答率が大きく伸びていることが確認できる．この2箇所では，カテゴリ数が4から3，あるいは3から2に減っている．区切り幅が広がることよりも，カテゴリ数の減少が正答率の伸びに大きく影響しているものと思われる．

3-2. ジャンルを考慮した発表年の推定

3-1.において，歌詞を分析することが発表年の推定に効果があることが確認できた．その際に，ジャンルを合わせて考慮することは有効であろうか．

流行歌には様々なジャンルが存在する．現在では，J-POP，K-POP，声優モノ，バラード，演歌，等々様々ある．しかし，今回の対象期間である70年にわたって長期間存在し，かつ対象データに含まれるものに限定すると，調査対象となり得るジャンルは限られる．

数あるジャンルの中で，演歌は最も息が長く，調査対象にも663曲中90曲含まれている．そこで，演歌に着目し，演歌と演歌以外に対象を分けて，それぞれについて発表年の推定を試みた．

ただし，演歌が年間トップ10にランクインしたのは1987年が最後であったので，対象データを1989年までのものに限定し，367曲（うち90曲が演歌）を対象に調査を行った．

なお，演歌であるかの分類については，筆者の判断に基づき，手作業で行った．

結果は，表3に示すとおりである．調査対象期間が40年間と短いため，区切り幅も最大10年に抑えている．

表3　ジャンルを考慮した発表年推定の結果

	全体	演歌	演歌以外
5年区切り	31.1%	23.3%	31.5%
8年区切り	45.8%	40.0%	48.9%
10年区切り	57.8%	47.8%	56.5%

この結果を見ると，演歌に関する発表年推定の正答率は，演歌以外のものに比べて，低い値を示している．演歌に関する正答率が低いことから，演歌を含む全体の正答率よりも演歌以外に限定した正答率のほうが高くなる．

この結果から，少なくとも1980年代までにトップ10入りした演歌については，歌詞に使われる語の発表年に伴う変化は，演歌以外に比べると小さかったということがわかる．他のジャンルに比べれば，いつの時代も演歌は変わらないということを示していると言える．

3-3. 発表時の景気動向の推定

世の中の状態を評価するための観点は様々ある．その中でも人々が頻繁に口にするのが，“景気の良し悪し”である．人が“景気”というとき，それが指すものの実態は明確ではない．つまり具体的な数値指標を基にしているわけではないが，日々の生活から得られる印象を一言にまとめて，「景気が良い」或いは「景気が悪い」という言葉を頻繁に口にする．

一方，流行歌が反映する世相も同様な“ザックリとした”ものと考えられ，両者の間には何らかの相関があると期待される．そこで，流行歌の歌詞を分析し，その歌が発表された時期が景気のいい時期だったか，或いは景気の悪い時期だったかの推定を試みる．

ここでは，景気の判断の基準として政府が発表する景気基準日付[8]（表4）を用いる．ここには，景気の波の山と谷が示されている．谷から山に至る期間を拡張期間と呼ばれ，この間は景気が良くなる．一方，山から他に至る期間は後退期間と呼ばれ，この間は景気が悪化する．

ここでは，対象期間を[8]をもとに拡張期間と後退期間に分け，歌詞からそのどちらに発表された歌であるかを推定する．

ただし，景気の拡張期間と後退期間が切り替わった瞬間に世間の雰囲気が一変するわけではなく，両者の間には“ズレ”があることが予想される．そこで，政府発表の日付を基準に，2か月単位で前後 4 ヶ月の範囲で景気の切り替わりの時期をずらして判定を繰返し，どうずらした場合に正答率が最も高くなるかを調査する．

政府による景気基準日付が発表されているのは，1951 年 6 月以降についてだけである．そこで，対象とする歌を1951年6月以降に発表された641曲に限定して調査を行った．

山と谷をずらす範囲を前後 4 ヶ月分までとしたのは、単一の期間の最短が第1循環の後退期間である1951年6月～同年10月の4ヶ月だったためである．

結果を表5に示す．処理に用いた期間を拡張期間終了と後退期間終了の時期に対してどれだけずらしたかの組み合わせごとに正答率を整理してある．

いずれも前後にずらすことなく推定した場合の正答率は67.4%であった．ランダムに選んでも50%の正答率が得られるので，それほど高い値とは言えない値である．

最も高い正答率が得られたのは，拡張期間終了から4ヶ月後ろに，後退期間終了から4ヶ月前にずらした場合で，85.0%の正答率であった．これは，50%に比べて高い正答率と言える．

表5から，拡張期間終了の区切りからは遅らせるほど，後退期間終了の区切りからは早めるほど，正答率が増加することが確認できる．このことは，

景気が下降線をたどり始めてもしばらくの間歌は好景気を反映し続けるのに対して，景気回復についてはその兆しを実際よりも早く映し始めると解釈できる．

全体として，景気の拡張期間を写した歌が好まれる傾向があるとも言えそうである．

表 5　期間終了との差の組み合わせと正答率

（単位：%）

		後退期間終了の区切りとの差(ヶ月)				
		-4	-2	0	2	4
拡張期間終了の区切りとの差(ヶ月)	-4	68.0	66.5	61.5	61.5	60.4
	-2	70.5	67.9	64.9	60.5	60.8
	0	75.5	71.8	67.4	63.2	61.5
	2	78.5	74.7	71.6	66.8	64.0
	4	85.0	81.0	78.0	73.6	69.1

4．「夢」の使われ方

今回研究対象とした楽曲の歌詞データにおいて，出現回数の多い名詞を調べた結果を表 6 に示す．出現回数，出現曲数ともに「夢」という語が最も多いことがわかる．

表 6　出現回数及び出現曲数の多い名詞

抽出語	出現回数	抽出語	出現楽曲数
夢	429	夢	258
愛	409	心	212
恋	390	恋	204
心	352	涙	203
涙	320	愛	200
人	317	胸	185
胸	267	人	182
風	247	風	155
空	224	手	141
男	217	空	139

「夢」は多義語の一つであり，デジタル大辞泉[9]によれば主な意味として次の五つが挙げられている．(2)と(3)はポジティブな意味合いであるのに対して，(4)と(5)はネガティブな意味合い，(1)はニュートラルな意味合いと言える．

(1) 睡眠中に，あたかも現実の経験であるかのように感じる一連の観念や心像
(2) 将来実現させたいと思っている事柄
(3) 現実からはなれた空想や楽しい考え
(4) 心の迷い
(5) はかないこと

1950 年代，1980 年代，2010 年代に発表された楽曲を対象として，歌詞中に登場する「夢」という語の使われ方を集計した結果を表 7 に示す．年代を追うごとに，ネガティブな意味合いで使われることが減り，逆にニュートラル或いはポジティブな意味合いで使われることが増えている．この結果は，[4]における『1997 年を境界に，暗い内容の楽曲が減少し，明るい内容の楽曲が増えた』との考察と合致する内容と言える．

表 7　「夢」の意味別登場回数

	※1	※2	※3	※4	※5
1950年代	5	9	14	3	14
1980年代	15	19	16	8	3
2010年代	20	29	24	4	1

※1：睡眠中に，あたかも現実の経験であるかのように感じる一連の観念や心像
※2：将来実現させたいと思っている事柄
※3：現実からはなれた空想や楽しい考え
※4：心の迷い
※5：はかないこと

5．まとめ，今後の課題

歌詞の分析を行うことで、楽曲の発表年の推定ができうること，ジャンルの推定については，（今回調査した範囲では）その効果があまり見られなかったこと，景気拡張期間を広げるような解釈が好まれるということが確認できた。

歌詞に登場する「夢」という単語の使われ方については，「将来実現させたい事柄」「現実離れの空想」というポジティブな意味合いでの使用頻度が増加している．

年代別のカテゴリ分類での分析において，実際よりも前の年代の特徴を持つ楽曲が多いと考えられる結果が得られた理由について，より詳細な検討が望まれる．

景気基準日付の拡張期間・後退期間の 2 種類のカテゴリに分類しての分析において，カテゴリ内のデータ数を揃えるなどの条件変更を行ったうえでの検討も望まれる．

また，年代や景気基準日付以外の分類基準での検討についても望まれる．

歌詞中に登場する特定の単語の使われ方について，「夢」以外の単語での検討も望まれる．

参考文献

[1] 福田慎一，照山博司.『マクロ経済学・入門』. 有斐閣アルマ, 1996.
[2] 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 論文集, 2010, Vol.2010, No.15, pp.195-202.
[3] 小林佳織，狩野恵里奈，鈴木崇史. 女性グループの歌詞の計量テキスト分析. https://www.anlp.jp/proceedings/annual_meeting/2013/pdf_dir/P1-5.pdf, (参照 2020-9-1).

[4] 大出彩, 松本文子, 金子貴昭. 流行歌から見る歌詞の年代別変化. じんもんこん 2013 論文集, 2013, Vol.2013, No.4, pp.103-110.
[5] “KH Coder”. http://khcoder.net/, (参照 2020-11-9).
[6] “年間シングルヒット曲”. https://entamedata.web.fc2.com/music/hit_top_single.html, (参照 2020-11-9).
[7] “うたまっぷ.com”. http://www.utamap.com/, (参照 2020-11-9).
[8] “景気基準日付”. https://www.esri.cao.go.jp/jp/stat/di/hiduke.html, (参照 2020-9-1).
[9] “夢とは - コトバンク”. https://kotobank.jp/word/%E5%A4%A2-145376 (参照 2020-11-9)