古典和歌集からのテキストマイニング

竹田, 正幸
九州大学大学院システム情報科学研究院情報理学部門

福田, 智子
純真女子短期大学

南里, 一郎
純真女子短期大学

https://doi.org/10.15017/4782068

出版情報：九州大学情報基盤センター年報. 1, pp.29-32, 2001-10. 九州大学情報基盤センター
バージョン：
権利関係：

情報基盤センター年報第1号（2001年）
九州大学情報基盤センター

古典和歌集からのテキストマイニング
Text Mining from Anthologies of Classical Japanese Poems

竹田 正幸† 福田 智子‡ 南里 一郎‡
Masayuki Takeda† Tomoko Fukuda‡ Ichiro Nanri‡

† … 九州大学大学院システム情報科学研究院情報理学部門
† … Department of Informatics, Kyushu University
‡ … 純真女子短期大学
‡ … Junshin Women's College

要旨　本論文では，文学作品，とりわけ古典和歌からのテキストマイニングに関する著者らの最近の研究の概略を述べる．歌集から特徴パターンを見出す方法と類似歌対を得る方法を示す．著者らの目的は，単に効率的なアルゴリズムを開発することではなく，和歌文学者の興味を引くような結果を得ることにある．実際，本研究で開発した類似歌抽出法により，いくつかの重要な文学的発見がもたらされた．

Abstract　This paper surbveys our recent studies of text mining from literary works, especially classical Japanese poems, Waka. We present methods for finding characteristic patterns in anthologies of Waka poems, as well as those for finding similar poem pairs. Our aim is to obtain good results that are of interest to Waka researchers, not just to develop efficient algorithms. We report successful results in finding patterns and similar poem pairs, some of which led to new discoveries.

1　まえがき

平成8年，約45万首の古典和歌を収めた『新編国歌大観』CD-ROM版が，角川書店より刊行された．著者らは，和歌文学研究の支援を目的として，このような和歌の集積に対するテキストマイニングの研究を行っている．本稿ではそれについて概観する．

本研究の特色は，以下のとおりである．

1. 自然言語処理を一切施さないこと．
2. 人手による品詞分解なども行わず，和歌を単なる仮名文字の連鎖とみなすこと．
3. 人手による従来研究を計算機になぞらせるのではなく，まったく別の視点と手法を用いてこれまで看過されていた事実の発見を目指すこと．

1は，統語処理や意味処理を行う解析器の精度が十分でないことに加え，テキストマイニングの結果が，解析器の用いている辞書や文法規則，解析器自体の性癖に依存してしまうことを避けたい，という理由による．2の理由としては，内省による品詞分解作業は，多大な労力を要することとともに，品詞分解を施すと，掛詞などをかえって見えにくくすることが挙げられる．また，3から，学習に必要なだけの訓練例は得られないものと考え，例からの学習 (learning by examples; e.g.[6]) の手法は適用しない．

なお，実験には，『新編国歌大観』CD-ROM版の句索引のデータファイルからもとの和歌を復元したものを利用した．これにより，すべて清音表記された仮名文字列のデータを得ることができる．

2　歌集からの特徴パターンの抽出

古典和歌における表現の分析は，これまで，もっぱら名詞や動詞を中心とする自立語に着目して行われてきた．これらの語は，表現の素材となり，「梅に鶯」「紅葉と鹿」のように，特定の組合せで用いられる．しか

し，自立語に偏した従来の研究は，片手落ちの誹りを免れない．なぜならば，自立語と自立語を連繋させ一首の和歌にまとめあげるという重要な役割を担う付属語(助詞・助動詞)が，ここではまったく度外視されているからである．

そこで，このような付属語重視の発想に基づいて，著者らは，付属語や用言の活用語尾などの作るパターンであるふし(節)を表現技法を特徴づけるモデルとして提案した[10, 14]．実例をあげておこう．

あききぬと/めにはさやかに/見えねども/
　風の音にぞ/おどろかれぬる　　『古今集』169番

これは，秋の到来を風の音で知るという，『古今和歌集』秋部冒頭の有名な歌である．この歌と，次の2首とを比べてみると，歌ことばの支配する表現世界がまったく異なるにもかかわらず，ある類似性が存在するのに気づく．

せきとめて/うぢのかはなみ/よせねども/
　つきにそひてぞ/心ゆきぬる　　『為仲集』97番
松しまの/あまのとまやは/しらねども/
　我が袖のみぞ/しをれわびぬる『後鳥羽院御集』1041番

すなわち，「ねども」「ぞ」「ぬる」という付属語もしくは付属語の列が，同じ順序で用いられているために，これらの歌が似ていると認識されるのである．このような構造を，ここでは「ふし」とよび，

*ねども*ぞ*ぬる*

のように表すことにする．

あるパターンに着目してその用例を調査し，意味のある結果が得られれば，それは研究成果となる．和歌のデータが機械可読化されていれば，パターンに合致する和歌をすべて取り出すことは容易である．だが，これまで，そのようなパターンは，研究者が任意に与えるしかなかった．もしそれを計算機を利用することにより自動的に抽出できれば，新たな発見への端緒となることも期待できる．

1つの歌集に表れるパターンの異なり数は数十万にものぼるため，そのすべてを研究者が吟味することは，現実には不可能である．そこで，その大量のパターンの中から「重要」と思われるものだけを，数百程度のオーダーで自動抽出することを考えたい．これが可能となれば，研究者はそれらのパターンを重点的に吟味することにより，有用な知見を得ることができよう．

文献[10, 14]では，最小記述長(MDL)原理[7]に基づいたパターン抽出法[2]を用いて，歌集からのふしの自動抽出を試みた．得られたふしの歌集ごとの相違は，歌人の個性や時代の好みを反映しており，研究者に非常に興味深い視点を与えるものであった．

3 類似歌発見

文献[11, 12]では，類似歌の半自動抽出について論じた．そこでは，意味的処理を一切行わず，また，単語という概念すら捨ててしまって，和歌を単なる仮名文字の連鎖とみなし，和歌間の共通部分文字列に着目して類似性を考える．このような観点で類似した和歌は，本歌取り，すなわち，特定の歌を踏まえて新しい歌を作る手法によるものであることが少なくない．また，本歌取りではなくとも，先行歌と同様の発想で詠まれた類想歌や，一首の歌が伝来の過程で本文の微妙な違いを生じた異伝歌であることもある．したがって，このような類似歌を見出す有効な方法が得られれば，和歌文学研究への大きな寄与が期待できる．

類似歌の抽出法として，和歌間の類似性指標を定義し，その指標の値の大きい和歌の対を人手により検証する，といった方式が考えられる．このような方式においては，成功の鍵は，類似性指標をいかに定義するかにかかっている．

有効な類似性指標が唯一存在するとは考えられない．むしろ，研究者の視点に応じて指標を自由に変更し，その都度，類似度の値の高い対を確認していく，というシナリオに沿った研究が有効であろう．そして，そのような指標の設計と変更は，場当たり的に行うのではなく，ある共通の土台の上で，見通しよく行うべきである．

そこで，まず，類似性指標のための統一的枠組みを導入した．この枠組みでは，指標を，パターン集合とパターンにスコアを与える関数との対によって表す．そして，二つの文字列間の類似度を，その共通パターンの最大スコアとして定義する．この枠組みは，

1. 代表的な非類似性指標である編集距離及びその変種をすべて表現でき，かつ，
2. 類似性が共通パターンとして陽に与えられるため指標を直感的に把握しやすい，

という利点をもつ．

次に，この枠組みのもとで，類似歌の半自動抽出に適した類似性指標を三つ提案した．第1の指標は，和歌を5-7-5-7-7の五句に分割し，句ごとに求めた類似度の総和を和歌間の類似度とするものである．句間の類似度は，パターン集合を正規パターン(regular pattern)[9]の集合とし，各パターンのスコアを，パ

ターン中の文字列の長さと個数に依存して定めるものである．また，第2の指標は，句に分割せずに，歌全体での共通部分文字列を求める．パターン集合としては，順序自由パターン (order-free pattern) の集合を用い，スコアはパターン中の文字列の長さに依存して与える．さらに，第3の指標は，パターン集合は第2の指標と同じであるが，パターンのスコアをパターンの生起頻度に依存して与えるものであり，稀少度が高いパターンを共通してもつ対ほど類似度は高くなる．

これら三つの指標を用いて，『古今集』と『新古今集』など，二つの和歌集の間のすべての対について類似度を算出し，類似歌の抽出を試みた．その結果，

- 類似度の高い対の多くは本歌取りであること．
- これまで指摘のなかった本歌取りの例を類似度の高いものとして拾うことができること．

が判明した．また，本歌取り以外にも，ある特定の詠歌状況下で用いられる表現や，伝来の過程で表現のバリエーションが生じた異伝歌，掛詞などの表現技巧が共通する歌などが抽出できた．特に，第3の指標を用いた場合には，その他の指標では類似度が下位になっていた歌の対が上位に浮上し，既知の常套表現をできる限り排除した，より緊密な類似性をもつ歌の対を得ることができた．

類似歌発見の研究の最終目標は，古典和歌における表現技法の系譜を明らかにすることである．本手法により，今まで見過ごされてきた表現の影響関係をいくつか見出すことができた．たとえば，親心の率直な吐露とのみ評価されてきた藤原兼輔の歌(『後撰集』1102番) が，清原深養父 (『古今集』585番) の骨組みを利用した，いわば「替え歌」であることを発見した [4]．これにより，古歌を踏まえた歌作りの一面が明らかになった．また，『為忠集』の成立年代について，これまで鎌倉中期頃かといわれてきたが，表現の授受関係から，実は室町時代であることを実証した [5]．これは，表現研究が歌集の成立年代推定にまで発展した例である．

4 歌集間における表現の差異の抽出

前章で述べたように類似歌抽出を行う一方で，著者らはさらに，以下のような場合の表現分析の方法を，新たに案出する必要性を感じる．すなわち，歌人Aと歌人Bが，親と子，師匠と弟子といった，近い間柄にあるときである．この場合，歌人Aの影響 (あるいは指導) を歌人Bが受け，類似歌を多く詠むことは容易に推測されるため，それらの多くは，既に研究者の手によって指摘されていることも少なくない．とすると，こういった歌人の家集間に見出される類似歌を，その上に追加していくよりむしろ，その差異を明らかにすることの方が，表現分析において，重要な観点になるであろう．もし，ある表現を，歌人Aが頻繁に使用した反面，歌人Bはほとんど (もしくは全く) 用いなかったとしたら，それがまた，和歌文学研究の糸口になるかもしれないのである．そこで，文献 [13] では，二人の歌人の家集から，その表現の頻度の違いをもとに，表現の差異を抽出することを目指した．これは，古典和歌の表現分析をする上で，著者らがこれまで行ってきた，表現の授受関係を見出すための類似歌抽出と，表裏一体をなすものである．

二つの歌集を入力とし，一方にはよく表れるが他方にはほとんど表れないパターンを得る問題は，テキストデータからの最適パターン発見問題 [8] において，扱うパターンを部分文字列パターンに制限したものと捉えることができる．部分文字列の個数は，本質的には入力長に比例するため，最適パターンは自明な線形時間アルゴリズムで得ることができる．パターンの「良さ」に関する統計的尺度としては，分類誤差，エントロピー，Gini 指標などがよく用いられる．しかし，この尺度をどう選んだとしても，得られたパターンがそのまま有用であることは期待できない．このため，上位のパターン群について，専門家の手によって吟味する作業が不可欠となる．部分文字列は，単語列の無意味な断片であることが多いため，作業負担は少なくない．その負担をいかにして軽減し，作業支援を行うかが，成功の鍵を握っているといってよい．

そこで，文献 [13] では，次のことを提案した．リスト中の冗長性を除くため，Σ^* 上の同値関係を導入することにより，テキストの集合 S の部分文字列全体からなる集合を同値類に分割する．すなわち，部分文字列のリストではなく，同値類のリストを作成するのである．この同値関係は，Blumer ら [1] によって定義されたもので，以下の性質をもつ．

- 各同値類は唯一の最長文字列を含む．同値類中の任意の元は，この最長文字列の部分文字列となっており，この最長文字列を代表元とみなす．
- 同じ同値類に属するすべての元は，テキスト集合 S において同じ頻度で生起し，したがって，「良さ」の値も同じである．
- 同値類の個数は，S 中の文字列長の総和に関し

て線形である.

各同値類の最長文字列である代表元を主要部分文字列とよぶ. 主要部分文字列とは, 直感的には, 部分文字列を, その頻度が変わらない限りにおいて左右両方向に延長して得られる部分文字列である. 左右両方向に延長するのではなく, どちらか一方にだけ延長したとしても, 得られる部分文字列の個数は, 入力長に比例する. 接尾辞木 (suffix tree)[3] は, 右方向への延長だけを考慮して得られる同値関係に基づくデータ構造であるといえる. DAWG(directed acyclic word graph)[3] は, 逆に, 左方向への延長だけを考慮した同値関係に基づく. 主要文字列の個数も, 入力長に比例するが, その個数を減らすことができる (後述の歌集については約 1/4 に減少した). また, ひとつの同値類に属する文字列の個数は, 最悪の場合, 代表元の長さ m の自乗に比例するが, 代表元と高々 m 個の極小元のみによって同値類すべての文字列をコンパクトに表現可能である.

提案した方法を用いることにより, 著者らは, 二つの家集からその差異を特徴づける文字列を抽出し, これを人手で調べていくことで, 特徴的表現の発見に成功した. すなわち, 西行の『山家集』と慈円の『拾玉集』, また, 藤原定家の『拾遺愚草』とその息子為家の『為家集』をそれぞれ比較し, その差異となる特徴表現を得たのである.

西行と慈円は, ともに歌僧であるとはいえ, 身分や境涯は極めて対照的であった. だが, 摂関家の子弟の身で僧籍に入り, 俗世間の権力と関わりを断てない立場の慈円は, 晩年の西行に, 隠遁の志を打ち明けるなど, 和歌だけではなく生き方にまで, 大きな影響を受けたことが知られている.

また, 為家は, 歌の家として名高い御子左家の嫡流として, 父定家から, 厳しい指導を受けた. 定家の歌の中には, 為家の手本になったものも, 少なからずある.

これら二組の歌集は, それぞれ, 類似歌が存する必然性を備えている. そのような歌集間において, 逆に, 表現の差異が抽出されたことで, それが, 個々の歌人のもつ, 見過ごされてきた一面の発見へとつながる可能性は, じゅうぶんに期待できる.

5 むすび

著者らのグループで行っている, 古典和歌データを対象にしたテキストマイニングの事例について述べた. 詳しくは, 文献を参照されたい.

参考文献

[1] A. Blumer, J. Blumer, D. Haussler, R. Mcconnell, and A. Ehrenfeucht. Complete inverted files for efficient text retrieval and analysis. *J. ACM*, **34**(3):578–595, 1987.

[2] A. Brāzma, E. Ukkonen, and J. Vilo. Discovering unbounded unions of regular pattern languages from positive examples, *Proc. 7th International Symposium on Algorithms and Computation (ISSAC'96)*, 95–104, 1996.

[3] M. Crochemore and W. Rytter. *Text Algorithms.* Oxford University Press, 1994.

[4] 福田智子. 「人の親の心は闇にあらねども」–藤原兼輔の歌再考– (投稿中).

[5] 福田智子. 『為忠集』再考, 和歌文学会第 46 回大会研究発表資料, 2000.

[6] P.D. Laird. *Learning from good and bad data*, Kluwer Academic Publishers, 1988.

[7] J. Rissanen. Modeling by the shortest data description, *Automatica*, **14**:465–471, 1978.

[8] S. Shimozono, H. Arimura, and S. Arikawa. Efficient discovery of optimal word-association patterns in large databases, *New Gener. Comput.*, **18**(1):49–60, 2000.

[9] T. Shinohara. Polynomial-time inference of pattern languages and its applications, *Proc. 7th IBM Sympo. Math. Found. Comp. Sci.*, 191–209, 1982.

[10] 竹田正幸, 福田智子, 南里一郎, 山崎真由美. 和歌データベースにおける特徴パターンの発見, 情報処理学会論文誌, **40**(3):783–795, 1999.

[11] 竹田正幸, 福田智子, 南里一郎, 山崎真由美, 玉利公一. 和歌データからの類似歌発見, 統計数理, **48**(2):289–310, 2000.

[12] M. Takeda, T. Fukuda, I. Nanri, M. Yamasaki, and K. Tamari. Discovering instances of poetic allusion from anthologies of classical Japanese poems, *Theor. Comput. Sci.*, to appear. Preliminary version in: Proc. DS'99 (LNAI 1721).

[13] M. Takeda, T. Matsumoto, T. Fukuda, and I. Nanri. Discovering characteristic expressions in literary works, *Theor. Comput. Sci.*, to appear. Preliminary version in: Proc. DS 2000 (LNAI 1967).

[14] M. Yamasaki, M. Takeda, T. Fukuda, and I. Nanri. Discovering characteristic patterns from collections of classical Japanese poems, *New Gener. Comput.*, **18**(1):61–73, 2000. Preliminary version in: Proc. DS'98 (LNAI 1532).