人文科学とコンピュータ　45−11
（2000．1．21）

和歌データからの類似歌発見のための類似性指標について

玉利 公一† 竹田 正幸† 福田 智子‡ 南里 一郎*

† 九州大学大学院システム情報科学研究科 ‡ 福岡女学院大学 * 純真女子短期大学

要旨．本歌取りとは，特定の歌をふまえて新しい歌を作る作歌手法をいう．本歌取りの半自動抽出を行うための有効な方法として，まず，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，いかに類似性指標を定義するかにかかっている．著者らは，以前に最長共通部分列 (LCS) の長さを用いた指標を変更することにより，新しい指標を提案し，それが本歌取りの半自動抽出に有効であることを示した．しかし，本歌取りには様々なふまえ方があるため，むしろ，研究者の視点に応じて指標を自由に変更でき，その都度，類似度の値の高い対を確認していく，というシナリオが有効であろう．本稿では，類似性指標を自由に設計するための共通の土俵となる統一的枠組みを提唱する．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の対によって表し，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．このため，このもとで設計したおのおのの指標は，直感的に把握しやすいという利点がある．本稿では，この枠組みのもとで，いくつかの指標を提示し，本歌取りの半自動抽出の観点から評価する．

On Similarity Measures for Finding Instances of Poetic Allusion in Anthologies of Classical Japanese Poems

Kouichi Tamari† Masayuki Takeda† Tomoko Fukuda‡ Ichiro Nanri*

† Department of Informatics, Kyushu University

‡ Fukuoka Jo Gakuin College * Junshin Women's Junior College

Abstract. In this paper we consider a problem of semi-automatically finding instances of poetic allusion in a collection of classical Japanese poems. The key to success is how to define a similarity measure. We give a unifying framework that captures the essence of many existing measures. It makes it easy to design new measures appropriate to the problem. In this paper, we propose new measures and evaluate them.

1 まえがき

本歌取りとは，特定の歌をふまえて新しい歌を作る作歌手法をいう．本歌取りの半自動抽出を行うための有効な方法として，まず，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，いかに類似性指標を定義するかにかかっている．著者らは，文献 [5] において，最長共通部分列 (LCS) の長さを用いた指標を変更することにより，新しい指標を提案し，それが本歌取りの半自動抽出に有効であることを示した．

本歌取りには様々なふまえ方があるため，有効な類似性指標が唯一存在するとは考えにくい．むしろ，研究者の視点に応じて指標を自由に変更でき，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．従って，類似性指標を自由に設計するための共通の土俵が必要である．

そこで本稿では，既存の多くの類似性指標を統一的に扱うことのできる枠組みを提唱する．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の対によって表し，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．このため，このもとで設計したおのおのの指標は，直感的に把握しやすいという利点がある．本稿では，この枠組みのもとで，文献 [5] の指標を含めたいくつかの指標を提示し，本歌取りの半自動抽出の観点から評価する．

2 研究の背景

古典和歌研究における表現技巧や作歌手法の分析は，これまで，もっぱら自立語に着目した研究がなされ，付属語にはあまり注意が払われていなかった．例えば，以下のような基本的な修辞技巧についても，自立語が担うことばの意味内容に主眼が置かれていたといえるであろう．

(1) 掛詞
(2) 枕詞
(3) 序詞
(4) 縁語

また，

(5) 本歌取り

と呼ばれる作歌手法を考察する際にも，自立語を中心とするきらいがある．特定の歌をもとにして新たな歌を作るという，この作歌法は，いわば「替え歌」作りである．もとの歌と「替え歌」とに共通する自立語は，比較的指摘しやすい．人手によって本歌取りを指摘する場合，自立語は，主題や情景等と直接結びつくために，記憶に残りやすいのである．その点，付属語およびその組合せは印象に残りにくく，数多くの和歌の中から，付属語に着目して有意の和歌を拾い出すのは，極めて困難である．

しかし，従来の自立語のみに偏した研究は，片手落ちの誹りを免れないであろう．先の本歌取りにしても，先行歌のどの部分をふまえるかは様々である．従って，自立語とともに，付属語の共通性をも視野に入れることで，より作歌の実際に近づくことができるのではないだろうか．また，5-7-5-7-7 という，音数律に制約のある和歌においては，共通する音を把握することにも，また大きな比重をおいてよい．

一例を挙げよう．

人のおやの/心はやみに/あらねども/
子を思ふ道に/まどひぬるかな/ (後撰集#1102)

この歌は，三十六歌仙のひとり，藤原兼輔 (877-933) の代表作で，子を思う親の心情をストレートに表現した，ほとんど無技巧な歌である，という共通理解を得てきた．ところがこの歌，実は，次の先行歌を踏まえて作られたものと見られるのである．

人を思ふ/心はかりに/あらねども/
くもゐにのみも/なきわたるかな (古今集#585)

この『古今集』歌と先の兼輔歌とを比べてみると，

ひと …/こころは … に/あらねども/
…/… かな

という一首の輪郭が共通する．そしてさらに，第二句の「やみ」(兼輔) と「かり」(古今集) は，ともに [a][i] という，母音が共通する語である．このように，兼輔歌は，共通する自立語こそ多くはないものの，『古今集』585 番歌と，きわめて高い類似性を示している．

すると，先の兼輔歌は，単に無技巧といって片づけられないことになる．彼の念頭には，上の『古今集』の恋歌があった．兼輔は，そこに詠まれた恋人への一途な愛情を，我が子に向けて「替え歌」に仕立てたのであろう．このような有名な歌でも，付属語や音の共通性を考慮することで，これまで忘れ去られていた一面が発見されることもある．

付属語を重視した作歌法を提唱した人物に，江戸時代の国学者，富士谷御杖がいる．彼は，初学者向け作歌指導書『和歌いれひも』[3] において，一首の古歌から「脚結 (あゆひ)」(助詞，助動詞，接尾辞) の類だけをそのまま取り，その間に自立語を入れていけば，「腰折れ」にならずに多くの歌が作れるという．例えば，以下のようにである．

はるのひの/ひかりにあたる/われなれど/
かしらのゆきと/なるぞわびしき (古今集#8)
↓
○○○○の/○○○に○○る/○○なれど/
○○○○○○と/○○ぞ○○○○
↓
やまかはの/きしねにあまる/みづなれど/
やよひをときと/おとぞかれゆく

このような発想に基づいて，著者らは，付属語のなすパターンである**ふし** (節) を表現技法を特徴づけるモデルとして提案した [2]．例えば，パターン「*ば*ざらましを*」は反実仮想という表現技法に対応する．文献 [2] では，最小記述長 (MDL) 原理に基づいた方法を用いて，歌集からの付属語パターン自動抽出を試みた．得られたパターンの歌集ごとの相違は，歌人の個性や時代の好みを反映し，研究者に非常に興味深い視点を提供する．

また，文献 [5] において著者らは，よく知られた文字列間の類似性指標のひとつである最小共通部分列 (LCS) 指標に変更を加え，和歌に適した類似性指標を提案した．この指標により，二つの歌集間のすべての組に対して類似度を算出し，それに従って組を降順に整列し，例えば，上位 100 組を人手によって検証する，といった方法で，類似歌の自動抽出を行い，これまで指摘のなかった本歌取りの例を指摘できることなどを示した．実際，上で兼輔の歌の本歌として掲げた古今集 585 番歌は，この手法を用い

て『古今集』と『後撰集』を比較した際に見い出したものである．

著者らは，このように，和歌に一切の自然言語処理を施さずに，和歌を単なる文字の連鎖，音の連鎖とみなした立場で研究を行っている．このようなアプローチの有効性は，上記二つの研究成果から十分確認することができた．

3 類似性指標の統一的枠組

この章では，はじめに，既存の類似性 (非類似性) 指標を概観し，次に，それらの指標を統一的に扱うことのできる統一的枠組みを導入する．この枠組みによる類似性の定義は，編集距離などの定義とは異なり，直感的に把握しやすいという利点がある．この枠組みによって，和歌のみならず，様々な応用分野において，問題に適した指標を，場当たり的でなく，見通しよく設計ができる．例えば，MIDI データにおいて音符列の類似性を扱う場合などにも有効である [9].

3.1 既存の指標

最も単純な類似性指標の一つとして，文字列間の最長共通部分列 (longest common subsequence; LCS) の長さがある．例えば，文字列 ACDEBA と ABDAC は，LCS として ADA と ABA の二つをもち，従って類似度は 3 である．図 1 (a) における文字列の alignment は，文字列 ADA が LCS であることに対応している．図において，LCS に関与している文字の対は，上下に並べて縦棒で結んであり，それ以外の文字は，それぞれ，空白と向き合うように並べている．したがって，LCS の長さは，縦棒を伴って並んだ文字の対の個数である，ということができる．

一方，よく知られた非類似性指標として，*Levenstein* 距離がある．Levenstein 距離は，編集距離 (edit distance) とも呼ばれ，一方の文字列を他方に変換するために必要な編集操作の回数の最小値として定義される．ここで用いる編集操作とは，文字の挿入 (insertion)・削除 (deletion)・置換 (substitution) の三つをいう．文字列 ACDEBA と ABDAC の Levenstein 距離は 4 であり，図 1 (b) はその様子を図示するものである．二つの文字列の第 2 の文字 C, B が縦棒なしで向き合っていることに注意しよう．このような対は置換操作に対応する．一方，空白記号とそれに向き合った文字との対は，文字の挿入もしくは削除操作に対応する．図 1 (c) の alignment は，もう一つの LCS である ABA を与えるが，Levenstein 距離を与えるものではない．実際，この並びでは，5 回の編集操作を必要とする．

類似性と非類似性とは，双対的な概念である．例えば，LCS 長指標は Levenstein 距離と密接な関係がある．実際，編集操作として挿入・削除操作のみを許した場合には，二つの文字列間の距離は，文字列長の和から LCS 長の 2 倍を引いた値に等しい．置換操作をも許した場合にはこのことは成立しない．

分子生物学におけるヌクレオチド配列やアミノ酸配列の文字列比較には，もう少し複雑な指標が用いられる．まず，スコア関数 δ を以下のように定義する．$\delta(a,b)$ は，文字 a を文字 b に置き換えるコストを表し，$\delta(a,\varepsilon)$ と $\delta(\varepsilon,b)$ は，文字 a の削除や文字 b の挿入にかかるコストをそれぞれ表す．二つの文字列間の距離は，一方を他方へ変換する際の最小コストとして定義される．この指標を一般化 Levenstein 距離 (generalized Levenstein distance) とよぶ．通常，この δ は，距離の公理を満たすように定義される．

分子生物学においては，これ以外の編集操作がしばしば用いられる．例えば，連続した文字列の一括挿入や一括削除などの操作である．このような操作に関わるコストは，ギャップペナルティ (*gap penalty*) とよばれ，その値は，ギャップの長さの関数として与えられることが多い．通常，このギャップ関数として，一次関数や凸関数が用いられる．編集操作としては，ほかに *swap*, *translocation*, *reversal* などがあるが，これらについては，文献 [1] などを参照されたい．

3.2 統一的枠組み

多くの指標に対して，二つの文字列間の類似性 (非類似性) は，その両方に合致するパターンの最大 (最小) スコアと捉えることができる．すると，指標間の違いは，

- 共通パターンの属するパターン集合 Π.
- Π の各パターンにスコアを与えるパターンスコア関数 Φ.

の二つということになる．例えば，パターン集合として正規パターン全体を用い，パターン中の文字の個数をそのパターンのスコアとすると，LCS 長指標が得られる．実際，文字列 ACDEBA と ABDAC は，共通パターン A*D*A* を含むが，これは三つの文字を含んでいる．

形式的な議論は以下の通りである．Σ をアルファベットとする．**パターン**とは Σ 上の言語の表現であり，パターン π の表現する言語を $L(\pi)$ で表す．パターン π が文字列 $w \in \Sigma^*$ に**合致する**とは，$w \in L(\pi)$ であるときをいう．パターン π が二つの文字列 $x, y \in \Sigma^*$ の**共通パターン**であるとは，π が両方に合致すること，すなわち，$x, y \in L(\pi)$ であることをいう．以上で，文字列間の類似性指標 (非類似性指標) を定義す

```
AC DEBA      ACDEBA      ACDEB A
|  |  |      | |  |      |    | |
A BD  AC     ABD  AC     A    BDAC
```

(a) (b) (c)

図 1: Alignments

るための準備が整った．Σ 上の文字列の**類似性指標**とは，以下を満たす対 $\langle \Pi, \Phi \rangle$ をいう．

- Π はパターンの集合であり，
- Φ は Π の各パターンに実数値を割り当てる関数で，パターンスコア関数とよぶ．

類似性指標 $\langle \Pi, \Phi \rangle$ のもとでの文字列 $x, y \in \Sigma^*$ の類似度 $\mathrm{SIM}_{\langle \Pi, \Phi \rangle}(x, y)$ を，次式で定義する．

$$\mathrm{SIM}_{\langle \Pi, \Phi \rangle}(x, y) = \max\{\Phi(\pi) | x, y \in L(\pi)\}. \quad (1)$$

非類似度の場合には，上式において，最大値でなく最小値をとる．

上では，パターンを，一般的に，アルファベット上のある言語を定義する表現 (description) としたが，パターンを文字とワイルドカードから成る列に制限しても，実に様々な類似性指標を扱うことができる．

形式的には，**ワイルドカード** (*wildcard*) は，Σ^* の 1 個以上の文字列と合致する表現である．ワイルドカード γ が合致する文字列全体の集合を $L(\gamma)$ で表す．明らかに，$\emptyset \subset L(\gamma) \subseteq \Sigma^*$ である．Δ をワイルドカードの集合とする．パターンを $\Sigma \cup \Delta$ 上の記号列に制限する．Σ の任意の記号 a に対して，$L(a) = \{a\}$ とする．パターン $\pi = \gamma_1 \cdots \gamma_m$ の言語 $L(\pi)$ を，言語 $L(\gamma_1), \ldots, L(\gamma_m)$ の連接として定義する．

ここでは，基本的なワイルドカードとして，表 1 に示した 4 種類を導入する．さらに，括弧 $[,]$ を導入し，ワイルドカード γ に対して新たなワイルドカード $[\gamma]$ を $L([\gamma]) = L(\gamma) \cup \{\varepsilon\}$ によって定義する．例えば，文字 $a \in \Sigma$ に対して，ワイルドカード $[a]$ は空文字列 ε と文字 a のいずれとも合致する．同様に，ワイルドカード $[\phi^{(n)}]$ は，空文字列 ε および長さ n の任意の文字列と合致する．

ワイルドカードの集合 $\Delta_1, \ldots, \Delta_5$ を以下のように定める．

$$\begin{aligned}
\Delta_1 &= \{*\}, \\
\Delta_2 &= \{\phi\}, \\
\Delta_3 &= \{\phi, [\phi]\}, \\
\Delta_4 &= \{[a] \mid a \in \Sigma\} \\
&\quad \cup \{\phi(a|b) \mid a, b \in \Sigma \text{かつ} a \neq b\}, \\
\Delta_5 &= \Delta_4 \cup \{[\phi^{(n)}] \mid n \geq 1\}.
\end{aligned}$$

さらに，$k = 1, \ldots, 5$ に対して，$\Pi_k = (\Sigma \cup \Delta_k)^*$ とする．このとき，主な指標は以下のように表すことができる．

例 1 $\Phi_1(\pi)$ をパターン $\pi \in \Pi_1$ における Σ の記号の生起回数とするとき，対 $\langle \Pi_1, \Phi_1 \rangle$ は LCS 長指標を与える．ここで，写像 $\Phi_1 : \Pi_1 \to \boldsymbol{R}$ は，$\Phi_1(a) = 1$ $(a \in \Sigma)$ と $\Phi_1(*) = 0$ で定まる準同型写像である．

例 2 $\Phi_2(\pi)$ をパターン $\pi \in \Pi_2$ における ϕ の生起回数とするとき，対 $\langle \Pi_2, \Phi_2 \rangle$ は Hamming 距離を与える．ここで，写像 $\Phi_2 : \Pi_2 \to \boldsymbol{R}$ は，$\Phi_2(a) = 0$ $(a \in \Sigma)$ と $\Phi_2(\phi) = 1$ で定まる準同型写像である．

例 3 $\Phi_3(\pi)$ をパターン $\pi \in \Pi_3$ における ϕ と $[\phi]$ の生起回数とするとき，対 $\langle \Pi_3, \Phi_3 \rangle$ は Levenstein 距離を与える．ここで，写像 $\Phi_3 : \Pi_3 \to \boldsymbol{R}$ は，$\Phi_3(a) = 0$ $(a \in \Sigma)$ と $\Phi_3(\phi) = \Phi_3([\phi]) = 1$ によって定まる準同型写像である．

例 4 $\Phi_4 : \Pi_4 \to \boldsymbol{R}$ を，$\Phi_4([a]) = \delta(a, \varepsilon)$, $\Phi_4(\phi(a|b)) = \delta(a, b)$, $\Phi_4(a) = 0$ $(a, b \in \Sigma)$ によって定まる準同型写像とするとき，対 $\langle \Pi_4, \Phi_4 \rangle$ は，3.1 節で述べた一般化 Levenstein 距離を与える．

例 5 $\Phi_5 : \Pi_5 \to \boldsymbol{R}$ を，$\Phi_5([\phi^{(n)}])$ が長さ n のギャップに対するペナルティとし，それ以外は例 4 の Φ_4 と同様に定めるとき，対 $\langle \Pi_5, \Phi_5 \rangle$ は 3.1 節で述べたギャップペナルティ付き一般化 Levenstein 距離を与える．

4 和歌に適した類似性指標

この章では，和歌の本歌取りの半自動抽出に適した類似性指標の設計について論じる．

4.1 句の順序の変化

本歌取りにおいては，先行歌の表現が少なからず用いられることになる．したがって，歌人は単なるイミテーションに堕してしまわないように，注意を払う必要があった．藤原定家は『近代秀歌』(1209)『詠歌大概』(1221) において，以下のように記している [7, 6].

表 1: ワイルドカード

$*$	:	Σ 上の任意の文字列と合致するワイルドカード;
ϕ	:	Σ の任意の文字と合致するワイルドカード;
$\phi^{(n)}$	:	Σ 上の長さ n の任意の文字列と合致するワイルドカード ($n \geq 1$);
$\phi(u_1\|\cdots\|u_k)$	:	Σ 上の文字列 $u_1, \ldots, u_k$ のいずれとも合致するワイルドカード.

- 古歌を取りて新しき歌を詠ずる，五句の内に三句に及ばば，頗る過分，珍し気なし．二句の上に三字，四字これを許す．
- 同じ言葉をもて古歌の詞を詠ずるは念なし．花をもて花を詠じ，月をもて月を詠ず．四季の歌をもて恋・雑の歌を詠じ，恋・雑の歌をして四季の歌を詠ず．かくの如き時は，古きを取るに難なし．
- 五七五の七五の字をさながら置きて，七々の字を同じく続けつれば，新しき歌に聞きなされぬ所ぞ侍る．

3 番目の項目によって，句の順序を入れ替えるように教えているため，われわれは，句の順序の変化を想定しなければならない．実際，次に示す本歌取りの例では，『古今集』147 番歌の初句，第二句，第四句に対して，『新古今集』216 番歌の初句，第四句，第二句が，それぞれ対応している．

例 6

ほととぎす ながなくさとの あまたあれば
猶うとまれぬ 思ふものから (古今集#147)

ほととぎす 猶うとまれぬ 心かな
ながなく里の よそのタぐれ (新古今集#216)

短歌形式の和歌は，五つの句から成るため，$5! = 120$ 通りの対応付けを考えなければならない．また，上の例では，3 対とも文字列として同一であったが，一般には数文字が異なる．そこで，120 通りの対応付けのおのおのについて，対応付けられた句の間での類似度の総和を求め，これを最大にするような対応付けを考える．その最大値を和歌と和歌の類似度と定義する．すると，次には，句ごとの類似度をどのように定義するかが問題となる．

4.2 句ごとの類似性指標

著者らは，文献 [5] において，まず，LCS 長に基づく指標を用いて実験を行い，その結果から，共通パターンにおける文字の連続性を考慮すべきであるとの着想に達し，これに基づく指標を提案した．以下の例をみてみよう．

例 7

山里は/冬ぞさびしさ/まさりける
人めも草も/かれぬと思へば (古今集#315)

やどさびて/人めも草も/かれぬれば
袖にぞのこる/秋のしら露 (拾玉集#3528)

「やまさとは」と「やとさひて」の「や」と「と」で 2 文字，「ふゆそさひしさ」と「あきのしらつゆ」の「し」で 1 文字，「まさりける」と「そてにそのこる」の「る」で 1 文字，それぞれ一致している．しかし，これらはほとんど無意味である．これに対し，「かれぬとおもへは」と「かれぬれは」の間での「かれぬ」「は」で 4 文字一致した，というのは，意味がある．このような文字の偶然の一致は，形態素解析を行わない限り避けられない問題であるが，文字が連続していれば，ある程度偶然の一致の可能性が低くなると考えられよう．このような観点から，文字の連続性を重視したスコア付けを行うことにした．

その指標は以下のようなものである．パターン中の連続文字列の長さに注目する．例えば，$\pi = *a*bc*d*$ においては，左から，$1, 2, 1$ である．正整数全体の集合 $\mathcal{N}$ から正実数全体の集合 $\mathcal{R}^+$ への写像 f を仮定し，パターンスコア関数の値を $\Phi(\pi) = f(1) + f(2) + f(1)$ のようにすることを考えよう．ここで，特に，$f(\ell) = \ell$ $(\forall \ell \in \mathcal{N})$ とすれば，$\Phi(\pi)$ は，π 中の文字の個数に一致する．文字が連続している場合に大きい値を与えるようにするためには，任意の正整数 n, m に対して

$$f(n+m) > f(n) + f(m) \qquad (2)$$

でなければならない．この条件を満たす f は無数に存在する．ここでは，$f(\ell)$ を ℓ の 1 次関数に限定し，

$$f(\ell) = \ell - s \qquad (0 < s < 1) \qquad (3)$$

とおいた．このパラメータ s としては，4.3 節で述べるように，本歌取りに関する正例・負例に対してある評価関数を最大にするものを選ぶことにする．

ここで，f として，2 次関数や，より高次の多項式関数を用いれば，パターン中の連続文字列の長さが長くなるに従って，類似度が急激に増加するようにできる．しかし，類似度は句ごとに計算するが，句の長さは高々 5 もしくは 7 であるから，問題となる連続文字列はあまり長くならず，顕著な効果は期待

できない．また，正例・負例として使用できるデータが少量かつ不完全なものであるため，学習すべきパラメータが多いと，過剰適合 (overfitting) の危険性があるため，上記の制限は妥当なものと考えられる．

4.3 類似性指標の評価

類似性指標のパラメータの学習，あるいは指標の良さの評価を行うためには，十分な数の本歌取りの例が必要である．しかし，例えば最もよく研究がなされている『古今集』と『新古今集』についてですら，本歌取りに関する完全なリストは存在しない．

そこで，慈円の『拾玉集』の一部を利用することを考えた．すなわち,『拾玉集』の 3,472 番から 3,571 番までの 100 首は,『古今集』の歌を踏まえて詠まれたものであり，題詞にその本歌が示してあるため，これを類似歌の例として用いるのである．

『拾玉集』の上述の 100 首とその本歌から成る 100 対を，類似歌の例，すなわち正例 (positive examples) とする．また，この 100 首と本歌以外の古今歌とから成る 9,900 対を，非類似歌の例，すなわち負例 (negative examples) とする．いま，

$$Succ_P(t) = \frac{\text{類似度が閾値 } t \text{ 以上の正例数}}{\text{正例数}},$$

$$Succ_N(t) = \frac{\text{類似度が閾値 } t \text{ 未満の負例数}}{\text{負例数}}$$

とおき，その相乗平均

$$Score = \sqrt{Succ_P(t) \cdot Succ_N(t)} \qquad (4)$$

をパラメータ学習のための評価関数として用いる．

句と句の類似性指標として文献 [5] の類似性指標を用いた場合には，最適なパラメータ s の値は，$s = 0.8 \sim 0.9$ 付近であり，閾値を変化させた際の最大値は

$$\sqrt{0.9600 \cdot 0.9608} = 0.9604$$

となった．この値は，LCS の長さを指標として用いた場合の値

$$\sqrt{0.9200 \cdot 0.9568} = 0.9382$$

と比べ，値が良いことがわかる．

しかし，これはあくまで，慈円の『拾玉集』の 100 首を用いた評価に過ぎず，特定の歌人に限定している点など，問題も多い．いずれにせよ，本歌取りに関して完全な正例・負例のデータが十分な量得られない以上，類似性指標の評価は，類似度の高いものが実際に本歌取りであるか，人手でチェックするほかない．和歌文学研究への貢献を考えるのであれば，むしろ，これまで看過されていた本歌取りの例を類似度の高いものとして指摘できるかが，最も重要な評価基準といえよう．

4.4 歌集間の類似歌抽出

『古今集』1,111 首と『新古今集』2,005 首の間の 220 万を超える組合せの各々について，類似度の値を計算し，その上位のものについて実際に本歌取りであるか検討した．類似度の値が 11 以上となる 73 対のうち 50 対が，代表的な注釈書 [4, 8] において，本歌取りとして指摘されていることがわかった．残りの 23 対は，一部の例外をのぞいて，本歌取りではないと考えられるものであった．

また，指標の値が 13 以上の 15 対については，13 対までが本歌取りとしての指摘があった．本歌取りでなかった 2 対を以下に示す．

例 8

すまのあまの/しほやく煙/風をいたみ/
おもはぬ方に/たなびきにけり (古今集#708)

しかのあまの/しほやく煙/かぜをいたみ/
立ちはのぼらで/山にたなびく (新古今集#1592)

例 9

春霞/たなびく山の/さくら花
見れどもあかぬ/君にもあるかな (古今集#684)

紫の/雲にもあらで/春霞
たなびく山の/かひはなにぞも (新古今集#1448)

両者は，文字列の共通性からは非常に類似している．特に，後者は,「はるがすみたなびくやまの」という和歌によくみられる慣用的表現を共通に含んでいるために類似度が高い値になっている．そこで共通する表現の生起確率を考慮すれば，類似度の値を小さくすることが可能と考えられる．

類似度が下がるにつれ，注釈書に指摘がないものが多くなるようであるが，その中には，本歌取りと考えてしかるべきものも含まれていた．例えば，以下の 1 対である．

例 10

あふ事を/ながらのはしの/ながらへて
こひ渡るまに/年ぞへにける (古今集#826)

ながらへて/猶君が代を/松山の
まつとせしまに/年ぞへにける (新古今集#1636)

上述の注釈書 [8] には，別の古今歌，すなわち，

かくしつつ/とにもかくにも/永らへて
君が八千代に/逢ふよしもがな (古今集#347)

を本歌として挙げてあるが，先に示した 826 番の古今歌も，併せて本歌とすべきものと考えられる．なお，この対の類似度は，11.5，全体の順位は 55 位であった．

5 指標の改善

本稿で提案した類似性指標のための統一的枠組みのもとで，具体的問題点に即して指標の変更を行う．

5.1 句ごとに対応をとる問題

次の2首をみてみよう.

例 11

ふるさとは/よしのの山し/ちかければ
ひと日もみ雪/ふらぬ日はなし (古今集#321)
みよしのは/山もかすみて/白雪の
ふりにしさとに/春はきにけり (新古今集#1)

この本歌取りの例では，先行歌の第二句に現れていた「よしの」「やま」の語句が，新古今歌では初句と第二句に分かれて現れていることがわかる．文献 [5] で提案した類似性指標は，句と句の対応のみを扱うために，このような場合をうまく扱うことができない．そこで，句ごとの対応を考えずに，31文字の中に共通して現れる部分文字列に着目する．この際，共通部分文字列の生起順位は問わない．

文字列 $u_1,\ldots,u_n \in \Sigma^+$ $(n \geq 1)$ に対して，パターン $\pi(u_1,\ldots,u_n)$ を，言語

$$\bigcup_{\sigma} L(*u_{\sigma(1)}*\cdots*u_{\sigma(n)}*)$$

の任意の文字列に合致するものとする．ここで，$\bigcup_\sigma$ は，$\{1,\ldots,n\}$ のすべての順列 σ に対する集合和を表す．また，パターン $\pi(u_1,\ldots,u_n)$ に対するスコアを，ある関数 f を用いて，$\sum_{i=1}^{n} f(|u_i|)$ と定める．

共通部分文字列について連続性を考慮しない場合，すなわち，関数 f として $f(n)=n$ とするならば，すべて1文字単位で考えてよいから，各文字の生起頻度をそれぞれの和歌で調べて，その小さい方の値を合計すれば，求める類似度が得られる．すなわち，文字 c の和歌 x における生起回数を $n_c(x)$ で表すとき，和歌 x,y の類似度は，$\sum_c \min\{n_c(x),n_c(y)\}$ で与えられる．しかし，一般の f については，文字 c ごとに

$$\frac{\max\{n_c(x),n_c(y)\}!}{|n_c(x)-n_c(y)|!}$$

通りの可能性があり，これを各文字ごとに掛け合わせたすべての場合を検討する必要がある．従って，計算量としては，入力長の指数時間となる．しかしながら，入力長は31文字程度で増加しないため，実際の計算時間が耐えられるものであるなら，問題はない．

ここでは，関数 f として，

$$f(1)=0 \qquad f(n)=n \quad (n>1)$$

となるものを用いた．この類似性指標に対して，4.3節で用いた評価関数の値を求めたところ，

$$\sqrt{0.9900\times 0.9507}=0.9702$$

となり，この値だけみると，改善されていることがわかる．

表2に，二つの指標についての類似度の度数分布を示す．括弧の中の数値は，前述の注釈書 [8, 4] に本歌取りとの指摘があった対の数を示している．その割合に関しては，文献 [5] の指標と比べ，特に顕著な違いはないようである．

この新しい指標での第1位の二対は，以下の通りである．

例 12

さむしろに/衣かたしき/こよひもや/
我をまつらむ/うぢのはしひめ (古今#689)
はしひめの/かたしき衣/さむしろに/
待つよむなしき/宇治の曙 (新古今#636)

例 13

花のちる/ことやわびしき/春霞
たつたの山の/うぐひすのこゑ (古今#108)
霞たつ/春の山辺に/さくら花/
あかずちるとや/鶯のなく (新古今#109)

古今集689番歌の結句の「うぢ」と「はしひめ」が，新古今集636番歌では，結句と初句とに分かれているが，これらに対応して類似度が計算されていることがわかる．もう一方の対についても同様である．文献 [5] の指標では，これらの対の類似度は，それぞれ，11.5，10.5であり，その順位は，55位，121位であった．

5.2 頻度を考慮した指標

4.4節でふれたように，表現の頻度を考慮することによって，和歌においてよく用いられる表現が共通している場合に類似度を下げることを考えよう．

パターンに与えるスコアをその稀少度 (rarity) に基づいて与える．S を Σ^+ の有限部分集合とし，S に属する文字列の類似度のみを考える．いま，$x,y \in S$ $(x \neq y)$ とする．$Pos=\{x,y\}$ を正例の集合，$Neg=S-Pos$ を負例の集合とする．正例のすべてに合致し，分類誤差 $|Neg\cap L(\pi)|$ を最小にする x,y の共通パターン π を考えよう．この場合，$|Neg\cap L(\pi)|$ を最小にすることは，$|S\cap L(\pi)|/|S|$ を最小にすることと等価であるから，S における生起確率が最も小さいパターン π を求めることになる．

ここではより単純に，以下のようにした．文字列 $u \in S$ の生起確率を

$$p(u)=\frac{|S\cap L(*u*)|}{|S|}$$

とおく．パターン集合として，5.1節で用いた集合を用い．パターン $\pi(u_1,\ldots,u_n)$ に対するスコアを

$$-\sum_{i=1}^{n}\log(p(u_i))$$

表 2: 類似度の度数分布

括弧内の数は，注釈書 [8, 4] に本歌取りとの指摘があった対の数を示す．

文献 [5] の指標			新しい指標		
類似度	度数	累積度数	類似度	度数	累積度数
16–17	2(1)	2(1)	21	2(2)	2(2)
15–16	1(1)	3(2)	20	2(1)	4(3)
14–15	4(4)	7(6)	19	4(4)	8(7)
13–14	8(6)	15(12)	18	7(5)	15(12)
12–13	26(18)	41(30)	17	11(9)	26(21)
11–12	32(20)	73(50)	16	22(11)	48(32)
10–11	77( )	150( )	15	54(20)	102(52)
9–10	137( )	287( )	14	110( )	212( )
8– 9	332( )	619( )	13	247( )	459( )
7– 8	1066( )	1685( )	12	608( )	1067( )
6– 7	3160( )	4845( )	11	1506( )	2573( )
5– 6	10089( )	14934( )	10	3534( )	6107( )
4– 5	35407( )	50341( )	9	7847( )	13954( )
3– 4	134145( )	184486( )	8	20744( )	34698( )
2– 3	433573( )	618059( )	7	30312( )	65010( )
1– 2	873904( )	1491963( )	6	104053( )	169063( )
0– 1	717547( )	2209510( )	5	70713( )	239776( )
			4	370354( )	610130( )
			3	75323( )	685453( )
			2	792725( )	1478178( )
			1	0( )	1478178( )
			0	731332( )	2209510( )

と定める．ただし，長さ 1 の文字列 u に対しては $\log_2(u) = 0$ とした．

以上のように頻度を考慮して定めた指標を用いて，『古今集』と『新古今集』の間の比較を行った．集合 S としては，二十一代集を用いた．現在，その結果を詳しく評価している段階であるが，5.1 節であげた古今集 684 番歌と新古今集 1448 番歌の類似度は相対的に下がり，全体の 93 位になった．頻度を考慮しない前節の指標では，16 位であったため，目的とした効果は得られたようである．

参考文献

[1] S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proc. 36th Ann. Sympo. Found. Comp. Sci.*, pp. 581–592. Springer-Verlag, 1995.

[2] M. Yamasaki, M. Takeda, T. Fukuda, and I. Nanri. Discovering characteristic patterns from collections of classical Japaneses poems. *New Generation Computing*, Vol. 18, No. 1, pp. 61–73, 2000. Also Proc. DS'98 (LNAI 1532).

[3] 富士谷御杖. 和歌いれひも. 三宅清（編）, 新編富士谷御杖全集, 第 5 巻. 思文閣出版, 1981.

[4] 久保田淳. 新古今和歌集. 新潮日本古典集成. 新潮社, 1979.

[5] 山崎真由美, 竹田正幸, 福田智子, 南里一郎. 和歌データベースからの類似歌の自動抽出. 情報処理学会「人文科学とコンピュータ」研究報告, Vol. 98, No. 97, pp. 57–64, 1998.

[6] 藤原定家. 詠歌大概. 久松潜一（編）, 歌論集, 中世の文学, 第 1 巻. 思文閣出版, 1971.

[7] 藤原定家. 近代秀歌. 久松潜一（編）, 歌論集, 中世の文学, 第 1 巻. 思文閣出版, 1971.

[8] 田中裕, 赤瀬信吾. 新古今和歌集. 新日本古典文学体系. 岩波書店, 1992.

[9] 門田隆史, 石野明, 竹田正幸, 松尾文碩. 音符列比較における類似性指標の評価. 第 59 回情報処理学会全国大会講演論文集 (2), pp. 17–18, 1999.