八代集「桜の花」歌における作者の分類

山元啓史　東京工業大学
ホドシチェク ボル　大阪大学

要旨

「よみ人しらず」の歌を他の和歌と比較し分類するために、歌同士の類似度を計算を試みる。しかし、歌データを対で比較することでクラスタリングが可能であるかどうかはよくわからない。そこで、Jaccard 係数, Dice 係数を用い、和歌の類似度計算を行い、相互関係を分析した。貫之、躬恒、遍昭、素性、小町、伊勢の 6 名の身元の明らかな作者の歌を対象とし、歴史的にも説明できる分類になっているかどうかを考察した。和歌の材料は「桜」歌に限定した。２首ずつの類似度関係のみから得られた結果では部分的な親疎関係がわかるだけで、すべての対の類似度を再帰的に計算した結果とは異なることがわかった。これらから、和歌の 1 首ずつそのものは限られたテキスト量ではあるが、できるだけ多くの作者の歌を使って再帰的に計算することによって、和歌の作者同士の相互関係を明らかにできる可能性が出てきた。今後、よみ人しらずの位置づけをはじめ、作者の相互関係の記述にも貢献できると考える。
キーワード: 和歌, 分類, よみ人しらず, Jaccard 係数, Dice 係数, 八代集

Classification of the poets of ‘cherry blossoms’ songs in the Hachidaishū

Hilofumi Yamamoto　Tokyo Institute of Technology
Bor Hodošček　Osaka University

Abstract

In this paper, we examine whether it is possible to determine the author of a song using a classification of classical Japanese poets based on text data. Six authors with clear attributions were chosen to test the effectiveness of the text classification, which also allows examining whether the resulting classification can be linked to historical evidence. We calculated the similarities between the poets using the Jaccard index and Dice coefficient, and visualized them, and observed their mutual relationships. We limited the poem material to “sakura” songs. We found that the affinity observed in one-to-one comparisons is not necessarily the same as the affinity obtained by recursive calculation. From these analyses, we found that clues could be obtained to clarify the mutual relationship between the poets even in the situations where only limited text was available. These results can contribute to the description of influence relationships between poets, including that between anonymous poems.
Keywords: classical Japanese poetry, classification, sakura songs, anonymous poet, Jaccard index, Dice coefficient, Hachidaishū

1 はじめに

本プロジェクトでは歌ことば辞書の開発のための基礎情報として作者の分類を行っている。和歌相互関係、作者相互関係は、古代の言語や文化（文芸思潮）を探るのに有益な情報である。和歌には作者のわからない「よみ人しらず」の歌があり、文学的にも優れた歌が多く含まれているが、残念ながら現状では、これらは分類上すべて不明とせざるをえない。

「よみ人しらず」とは、和歌の作者が不明であることを表すことばで、勅撰和歌集などによく見られる (小久保 2014)。本当に作者が不明の場合、後に明らかになったが、そのままの場合、当初からわかっているが、明らかにしなかった場合など、さまざまである（前掲）。「よみ人しらず」の歌を賞賛する研究者もおり、特に『古今集』の「よみ人しらず」は、後の二十代集の多くの作の本歌になっており、研究に値する要素が含まれている (塚本 1995)。「夕暮は雲のはたてに物ぞ思ふあまつ空なる人を恋ふとて（古今集巻恋歌一よみ人しらず: 484）」は小野小町作としても十分に通ると評される（前掲）。また「郭公鳴やさ月のあやめ草あやめもしらぬ恋もする哉（古今集, 恋歌一, よみ人しらず: 469）」は「うちしめりあやめそかほる郭公なくやさ月の雨の夕くれ (新古今集, 夏歌, 藤原良経:220)」を想起すると言われる（前掲）。これらのことから、よみ人しらずの歌が、ある作者の歌に似ていることが、作者の特定に至らなくても、どういうジャンルに分類されるのかを確かめるだけでも大きな意義がある。

「よみ人しらず」だけでなく、詠まれた時期がはっきりしない歌もある。また、作者はわかっていても、生没年が不明であることもあり、人手で分類するには考慮すべき要素が多い。和歌の 1 首ずつに含まれる和歌の要素（単語、語順、語種、音韻など）によって、作者の推定できるかどうかはよくわからない。31 文字限定のテキスト量で可能かどうかもわからない。作者をグループ分けするとは、グラフ理論でいうコミュニティの分割 (Fortunato 2010) にあたる。歴史的に得られた事実（生没年、共同作業者名）だけでなく、コミュニティの数理的分析を通して互いに影響しあっているかどうか、コミュニティが重複するかどうか、などの基礎的なデータも得られる。このような作者の分類は当時の歌風を客観的に分析する上で重要である。そこで、テキストのみの情報を用いて、作者のコミュニティを割り出せるかどうかを確かめ、テキストの要素の一致率から、2 者間の類似度を計算し、作者が分類できるかどうかを検討する。手法の可能性を検討するため、本稿では「桜の歌」についてのみ報告する。

2 方法

目的は、類似度計算によって「よみ人知らず」の歌を分類できるか、を確かめることである。作者名がわかったとしても、身元のよくわからない作者の歌は使えない。作風に関する十分な研究成果が活用できなければ、あるコミュニティに分類されたとしても考察できない。歌人論・歌論が十分に記述・考察されている実績ある作者の歌を使って類似度の計算を行う。本研究では「桜」もしくは「桜」を意味する「花」を詠んだ和歌を使って、和歌の分類を試みる。作者として、紀貫之、凡河内躬恒、遍昭、素性、小野小町、伊勢の 6 名を選んだ。「よみ人知らず」の歌は 1 人の作者が詠んだ 1 首であるのが前提である。それぞれの作者から 1 名から 1 首ずつ、同じ部立（四季や恋などの部立）の和歌を選んだ。ただし、小町は明示的に「桜」とは言わず「花」として桜を詠んでおり、注釈、部立から「桜の花」を示す歌を選んだ。

本研究では、和歌に含まれる要素の集合として、語を比較することとする。和歌を単位分割し、歌毎の用語集合の類似度を、Jaccard 係数 (Jaccard 1912) を用いて計算し、計算方法の違いを見る。

和歌は 1 首ずつ 6 作者を比較計算した。テキスト量が少なく類似度計算が安定しない場合も考えられるので、桜に関する 6 名それぞれ作者の和歌を 2 首ずつ集め、2 種類の類似度表を作成した。類似度の値を用いて、ネットワークによって可視化した。

2.1 類似度の計算

類似度計算には、ピアソンの積率相関係数, Cosine 類似度, Euclid 距離, Jaccard 係数, Dice 係数, Simpton 係数などがあるが、後半の 3 者はいずれも手計算で求められる。本研究においては、Jaccard 係数と Dice 係数について検討する（図 1）。材料は短歌に限定したため、比較する 2 つのテキストはいずれも 31 文字である。2 首の類似度を求める際、片方のテキストが長いということはない。ただし、要素の数に AB 間で食い違いがある場合には、Jaccard 係数では類似度は低くなる。和歌には問題とないとは考えられそうではあるが、差集合の要素数の影響、すなわちテキスト A とテキスト B との

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

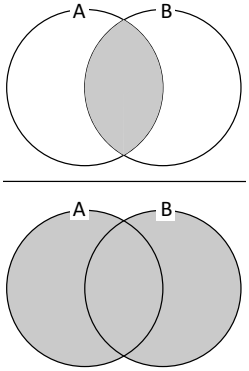

図 1: Jaccard 係数（式 J）と Dice 係数（式 D）による計算方法。短歌はテキスト長がほぼ 31 文字に限定されるため、A, B の要素数の食い違いはかなり少ないが、ないわけではない。任意２首の類似度を総当りで計算する。ただし、トピックが変わると類似度も低くなるため、本研究では「桜・花」歌に限定し、方法論の検討を行う。

要素の差（すなわち語数の差）があれば、自ずと分母よりも分子の方が値が小さくなってしまう。まったく単語数もいっしょだということは保証されないので、この問題を解消するために考案された Dice 係数についても検討する。

2.2 要素一致の方法

一致率は内容に依存するので、内容語を削除し、２文章の一致率を計算し、テキストの類似性を求めるのがよい。しかしながら、和歌においては、作者の内容よって使われる語の傾向も重要な要素である。また、古語の場合は、内容語であるか機能語であるかは、実際には明らかに区別できないこと、内容語が削除されると、計算に用いる要素がなくなってしまう場合も考えられることから、機能語だけでなく内容語も分析に用いる。ただし、内容が異なれば、類似しないのは当然なので、同じトピックの歌を選ぶ。和歌は部立によるトピックがだいたいの内容の分類として扱ってもよい。また部立の四季に関わる歌の並び方は、季節の移り変わりに応じているので、歌の並びによってだいたい同じような内容が配置されている。

自立語については、データ作成時に単語の正規化を行っていたため、表記において、同じ語であっても和歌に出現する単語が一致するか一致しないかは問題にならなかった。非自立語（助詞、助動詞など）については、助詞は表記どおり、助動詞は意味区分（「む」「らん」「まし」なら「推量」）で一致するものとした。小町が明らかに桜を意味する花を詠んだ歌が 2 首であったので、他の 5 作者についても国歌大観の歌番号の近い 2 首を選んだ（表 2.3）。ただし、「よみ人しらず」の歌は 1 首ずつであるので、1 首対 1 首で類似度を計算した。

2.3 材料: 八代集データベースの内容

材料として八代集データベースを用いる（表 1）。八代集データベースは、筆者らが 2010 年より科学研究費助成金を得て開発してきた。用語の単位情報（単位分割・品詞情報・作者名）とその正規化情報を備えた勅撰和歌集のデータベースである。元本となったデータベースは、新編国歌大観 CD-ROM 版の二十一代集に相当するデータ (新編国歌大観編集委員会 1996)、新日本古典文学大系本二十一代集に相当する書籍を参照し、正規化と単位分割を行った。作者と歌番号の紐づけは、国文学研究資料館編集二十一代集データベース (中村他 1999) にある作者タグを利用した。単語にあたる各単位は、国立国語研究所の分類語彙表に準ずる分類番号を施し、同表にない固有名詞（人名、地名、表外語彙）は新規番号をあて、単語検索と一致が計算できるようにした。

3 結果

表 3 は Jaccard 係数、表 4 は Dice 係数による類似度計算の結果である。素性と伊勢が最も近く（Jaccard .23; Dice .19）、Jaccard 係数では貫之・躬恒と小町（.06）、Dice 係数では貫之と小町が最も遠かった（.06）。類似度の値を重みとして、Graphviz の dot (Ellson et al. 2004) で描画したところ、躬恒と遍昭の間が遍昭と素性の間より近いのが少々異なるが、ほぼ貫之と躬恒、遍昭と素性、小町と伊勢にわかれた（図 2）。

表 1: 八代集収録の歌集一覧: 歌数は新編国歌大観による。

No.	名称	勅/院宣	成立	撰者	首
1	古今	醍醐天皇	905 頃	紀友則・紀貫之・凡河内躬恒・壬生忠岑	1100
2	後撰	村上天皇	951 頃	清原元輔・紀時文・大中臣能宣・源順・坂上望城	1425
3	拾遺	花山院	1007 頃	花山院	1351
4	後拾遺	白河天皇	1086	藤原通俊	1218
5	金葉	白河院	1125 頃	源俊頼	665
6	詞花	崇徳院	1151 頃	藤原顕輔	415
7	千載	後白河院	1188	藤原俊成	1288
8	新古今	後鳥羽院	1205	源通具・藤原有家・藤原定家・藤原家隆・藤原雅経・寂蓮	1978

表 2: 6 名の作者桜（花）の歌。? は生没不明。小町は桜を詠んでいないが桜を意味する歌を選んだ。

No	作者名 生没年	歌番号	歌
1a.	紀貫之	古今 049	今年より／春しりそむる／桜花／ちるといふことは／ならはさらなむ
b.	87?–945	古今 058	たれしかも／とめておりつる／春霞／たちかくすらん／山の桜を
2a.	凡河内躬恒	古今 086	雪とのみ／ふるたにあるを／桜花／いかにちれとか／風の吹らむ
b.	85?–92?	古今 358	山たかみ／雲ゐにみゆる／桜花／心のゆきて／おらぬ日そなき
3a.	遍昭法師	古今 394	山風に／桜吹まき／みたれなん／花のまきれに／立とまるへく
b.	816–890	古今 091	花の色は／霞にこめて／みせすとも／かをたにぬすめ／春の山かせ
4a.	素性法師	古今 055	見てのみや／人にかたらん／桜花／てことにおりて／いへつとにせむ
b.	???–910	古今 056	見わたせは／柳桜を／こきまぜて／都そ春の／錦なりける
5a.	小野小町	古今 113	花の色は／うつりにけりな／いたつらに／我身世にふる／なかめせしまに
b.	???–???	古今 797	いろみえて／うつろふ物は／世中の／人の心の／花にそありける
6a.	伊勢	古今 061	桜花／春くははれる／としたにも／人の心に／あかれやはせぬ
b.	872–938	古今 068	みるひとも／なき山里の／桜はな／ほかのちりなん／のちそさかまし

4 考察

いずれの計算方法においても、局所的には素性と伊勢の距離が最も近く、貫之・躬恒と小町の距離が最も遠いことがわかった（表 3）。全体的には、古今集編者の２名（紀貫之、凡河内躬恒）、僧侶（親子）の２名（遍昭、素性）、女性の２名（小野小町、伊勢）が比較的近い距離にあることがわかった（図 2）。和歌１対１で見ると素性と伊勢の歌が近く、貫之・躬恒と小町が遠いが、１対多で見ると部分部分で見た類似度で得られた距離とは異なることがわかった。Dice 係数による結果は、個々の類似度の値は Jaccard 係数とは異なるが、グラフ描画が示す相互関係には違いは見られなかったのは、興味深い結果である。グラフによる描画は、類似度表による局所的な類似関係からは得られない関係が見られ、いわば直感的である。反復再計算することにより、ノードとノードを結びつけるエッジの重みを類似度係数による値だけでなく引き合う力から総合的な図が呈示された。あるノード（素性）を一番強く引っ張っているノード（伊勢）は、バネ（スプリング）でできたエッジで引っ張っているため、ノード（素性）を引っ張るバネが他に多数あれば、伊勢が最も強く引っ張っていても、他のノードに引かれる力の方が総合的に強ければ、ノード（伊勢）とは遠ざかる。これは、バネ（スプリング）といった物理的なもののアナロジーに基づいているので、結果を予測したり、理解するのが簡単で、一般的には力学モデルと呼ばれている (Kamada and Kawai 1989)。これは、たとえばデンドログラム（最も近いものをまず結びつける作図法）にはない特徴である。１首あたりのテキスト量が少なく、１首に含まれる要素の偶然の一致がないわけではないが、多数の歌との類似度と比較されることにより、偶然性が少なくなる。均一的に整備されたテキストであれば、テキスト量は多くなくても、比較する対が多ければ、反復計算により漸近的なノード接続が行われ、このような図が得られたと考えられる。「桜」というよく詠まれる題材なので、安定した結果が得られたとも考えられるが、むしろ多少類似度にゆれがあっても、総当りで反復比較したので、この結果が得られたと考えた方がよさそうだ。和歌の要素だけで作者の分類を実施することが可能かどうかについてはさらにノード相互の接続関係から、作者の支配集合問題について分析することも考えるべきである。その上で、和歌テ

表 3: 桜歌による 6 作者の類似度表（Jaccard 係数）

No.	作者名	貫之	躬恒	遍昭	素性	小町	伊勢
1.	紀貫之	—	.19	.12	.12	.06	.14
2.	凡河内躬恒		—	.20	.15	.06	.17
3.	遍昭法師			—	.22	.10	.14
4.	素性法師				—	.10	.23
5.	小野小町					—	.16
6.	伊勢						—

表 4: 桜歌による 6 作者の類似度表（Dice 係数）

No.	作者名	貫之	躬恒	遍昭	素性	小町	伊勢
1.	紀貫之	—	.16	.10	.10	.06	.12
2.	凡河内躬恒		—	.16	.13	.09	.14
3.	遍照法師			—	.18	.09	.13
4.	素性法師				—	.09	.19
5.	小野小町					—	.14
6.	伊勢						—

キストの要素だけで歌・作者の分類が可能ということになるのであれば、限られたテキストの分類問題の解決に貢献できよう。

5 おわりに

本稿では、和歌の分類課題を歌データのみから Jaccard 係数および Dice 係数を用いて類似度計算を行い、グラフでその関係を一瞥した。いずれの計算においても、6 作者の分類は、局所的には、素性と伊勢が最も近く、貫之と小町が最も遠くなった。しかし、全体をグラフで描くと、古今集編者の２名（貫之、躬恒）、僧侶・親子の２名（遍昭、素性）、女性作者２名（小町、伊勢）の３つのグループにわかれた。局所的で得られた結果は、再帰的に計算して得られた結果と必ずしも同じでないことがわかった。歌 1 首それのみでは限られたデータではあるが、総当りで相互関係を再帰的に計算・分類すれば、当時の作者の相互関係を明らかにできる。これにより「よみ人しらず」の歌としての位置づけられている歌の分類や作者相互の影響関係の記述に貢献すると考えている。

参考文献

Ellson, J., E. R. Gansner, E. Koutsofios, S. C. North, and Gordon W. (2004) “Graphviz and Dynagraph: Static and Dynamic Graph Drawing Tools”, in M. Jünger and P. Mutzel eds. *Graph Drawing Software*, Berlin Heidelberg New York: Springer-Verlag, pp. 127–148.

Fortunato, Santo (2010) “Community detection in graphs”, *Physics Reports*, Vol. 486, No. 3, pp. 75 – 174. コミュニティの発見.

Jaccard, Paul (1912) “The distribution of the flora in the alpine zone”, *New Phytologist*, Vol. 11, p. 3750.

Kamada, Tomihisa and Satoru Kawai (1989) “An algorithm for drawing general undirected graphs”, *Information Processing Letters (Elsevier)*, Vol. 31, No. 1, p. 715.

小久保崇明 (2014) 『学研全訳古語辞典』, 学研教育出版.

中村康夫・立川美彦・杉田まゆ子 (1999) 『国文学研究資料館データベース古典コレクション 『二十一代集』(正保版本) CD-ROM』, 岩波書店, 東京.

新編国歌大観編集委員会（編） (1996) 『CDROM 版新編国歌大観』, 角川書店.

塚本邦雄 (1995) 「かなしき仮名序 —インタビュウ 塚本邦雄氏に聞く—」,『国文学 古今和歌集 —いま何が問題か学燈社』, 第 40 巻, 第 10 号, 16–25.

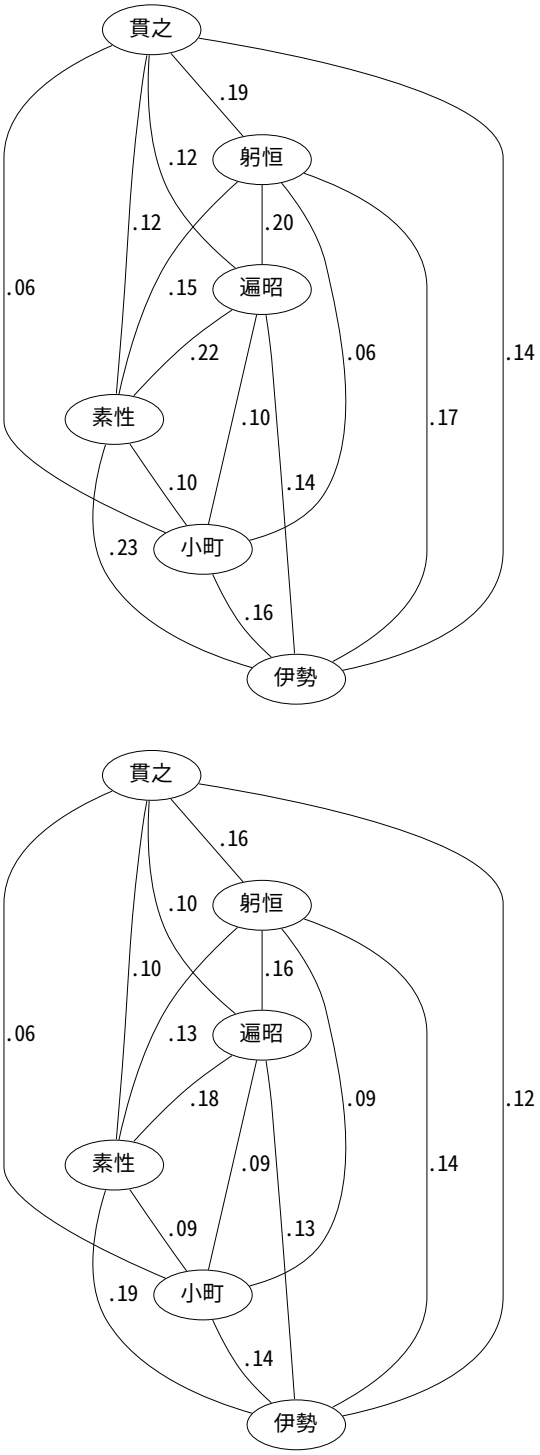

図 2: 作者 6 名間の類似度によるグラフ出力（上: Jaccard 係数; 下: Dice 係数）