

ブーリアン演算による歌ことばモデルの解析

山元 啓史

東京工業大学大学院社会理工学研究科

要旨

八代集 (905 年頃～1205 年) の和歌 (約 9500 首) を対象にグラフによる歌ことばのネットワークモデルを作成し、分析を行っている。語彙はノードとエッジの集合であり、それらで構成されるネットワークである。この集合に対して、和・差・積を求め、八代集における歌ことばの変遷を分析する。

キーワード：和歌、ブーリアン演算、ネットワーク、八代集、語彙、日本語史

An Analysis of the Models of Classical Japanese Poetic Vocabulary using Boolean Operation

Hilofumi Yamamoto

Graduate School of Decision Science and Technology, Tokyo Institute of Technology

Abstract

We have been analyzing the transitions of meanings of Japanese words using the network models of classical Japanese poetic vocabulary in the the *Hachidaishū* (ca. 905–1205). Vocabulary can be expressed as a class of nodes and edges, networks, which allow us to operate them mathematically. This paper addresses the analysis of network structures of classical Japanese poetic words using boolean operation: union, intersection, subtraction.

Keywords: classical Japanese poetry dictionary compilation morphological parser, the *Hachidaishū*, boolean operation

1 はじめに

語彙とは「語の集まり」のことであって、数えられる個々の語のことではない¹。本研究は「語の集まり」を集合とし、ブーリアン演算を用いて分析する方法について述べる。

語彙研究には「単語の離散的な集まり」として、単語の計量分析を主とする研究と、「組織的なまとまり」として単語と単語の類縁関係を分析する研究がある [24, 3]。語彙を「組織的なまとまり」として捉

える研究としては、日本語ソーラスのひとつである分類語彙表 [4] を基準に分類カテゴリ別に語彙の出現頻度を計算する手法が多く報告されている (たとえば, [12, 13, 23, 1, 16] など)。ただし、語相互の結びつきや依存関係に関する研究はあまり報告されていない²






筆者はこれまでに和歌用語を中心にグラフ表現を用いた語彙の分析を行ってきた [17, 18, 19, 21, 20]。これらの研究では、一首に共に出現する 2 語のパ

¹したがって、語彙数ではなく語彙量といい、英語でも vocabulary は uncountable である [11, 序論]。

²語という概念を用いず、n-gram 統計 (任意数の文字列長の統計量) を用いて歌ことばのジェンダー (男ことば、女ことば) を明らかにした研究はある [5, 6]。

ターンを1単位として、その集合をネットワークで表現し、分析を進めている。1語ではさまざまに解釈される語の意味も2語で分析すれば、その2語の示す文脈が想像しやすくなる。また、それらをグラフで示すことにより、鳥瞰図のように「語の集まり」が一瞥できる利点もある³。このようなグラフ図形は数理的表現で、論理和、論理差、論理積などのブーリアン演算を施すことができる。またその数理的性質をそのまま語彙研究に応用することができる。本論は、この点に注目し、ネットワーク中に見られる語彙の構成要素や依存関係の分析を示すことを通して、語彙研究の枠組みを提案するものである。

以下に用語Aと用語Bを中心とする2つの語彙ネットワークに対するブーリアン演算の種類を整理する。

1.  統合／論理和: AとBの2つのネットワークの統合。
2.  交差／論理積: AとBの2つのネットワークに共通して出現したもの。2つのネットワークが共有している語相互のつながり方と語相互の接続の量を視覚的に示す。
3.  差分A／論理差: ABの論理和から用語Bのネットワークを差し引いたもの。用語Bを排除し、用語Aにのみ関わる語彙を抽出する。
4.  差分B／論理差: 上記の逆。
5.  排他／否定論理積: 統合から交差を排除したもの。用語Aと用語Bの相違を強調する。

上記の演算を用い、a. 語の比較(類似する2語の具体的な相違の分析)、b. 時間の比較(語の2時代における比較)、c. 作者の比較(2名の作者の比較や性別による相違)、などが分析できると考えている。本稿では、aとbについて報告する。

³語彙を空間的に分析する方法はグラフではないが、マトリックスやデンドログラムを用いて語と語の相互関係を計算する研究は以前よりあった(主に[8, 9, 10]など)。

2 方法

材料は、国文学研究資料館編集正保本版「八代集」(古今集、後撰集、拾遺集、後拾遺集、金葉集、詞花集、千載集、新古今集)収録のすべての和歌9503首を用いる。和歌テキストは新編国歌大観の番号を付けた上でファイルにセーブした。それぞれの和歌テキストは、古文品詞タグ付けシステム kh [22]で単位分割し、品詞タグを付けた。分割の単位は国立国語研究所β単位にしたがった。単位分割だけでは、異表記同義語の問題があるので、それぞれの語を t2c⁴を使って、シソーラスコードに変換した。

モデルはあらかじめ出現する個々の語について idf [14, 15]⁵を計算し、次に共出現パターン(テキストに共に出現する任意2語の組み合わせ)を生成し、先程の idf 値とパターンの頻度を使って、各パターンの重みを計算して作成する。共出現パターンは単なる2語の組み合わせリストではあるが、共出現パターンで描画されたグラフには、もとの文にある文脈が含まれることがわかっている[17]。その点が単語リストによる頻度集計と異なる。

すべてのパターンを描くとグラフは真っ黒な塊になってしまうので、各パターンがそのテキスト群において、どの程度重要なパターンであるのかを評価し、重要なパターンから描き出す手続きが必要となる。そこで、テキスト群(d)において任意の1語(t)が特徴的であるかを評価する式 $tfidf(1)$ [7]を拡張し、任意の2語のパターン(t_1, t_2)がどの程度特徴的であるかを評価する式(2)を用い、パターンの重み(cw)を計算する。

$$w(t, d) = (1 + \log tf(t, d)) \cdot idf(t) \quad (1)$$

$$cw(t_1, t_2, d) = (1 + \log ctf(t_1, t_2, d)) \cdot cidf(t_1, t_2) \quad (2)$$

$$cidf(t_1, t_2) = \sqrt{idf(t_1) \cdot idf(t_2)} \quad (3)$$

ただし、(2)の前半は t_1 と t_2 の2語が共出現した時のテキストの数。(2)の後半 $cidf(t_1, t_2)$ は、(1)の $idf(t)$ を拡張し、2語の idf 値の幾何平均(3)としたものである。以上の方法で得られた cw 値を相互に比較できるよう、一旦標準得点に変換し、正規化を行い、1σ以上の共出現パターンを Graphviz (<http://www.graphviz.org/>)で描いた。

⁴t2c: Token to Code, 自作。単位切りした語を入力すると分類語彙基準のシソーラス体系コードを返すプログラム。

⁵ idf はある特定のテキストにしか出現しない語か、どんなテキストにも出現する語なのかを示す値。 $idf(t) = \log N/df(t)$ ただし、 N はすべての資料の数、 $df(t)$ は、語 t の出現する資料の数。

3 結果

3種類の演算の結果を示す。はじめに（これは厳密にはブーリアン演算ではないが）コアノード（分析する用語）を全体の集合より削除し、その余りの集合を分析する方法、つぎに、2語の集合を統合した時の論理積（交差）をグレーで示し、2語の近さを分析する方法、最後に、2語の関係を時代を隔てて分析する方法について述べる。

3.1 コアノードの削除

削除はプルーニング（枝の刈り込み）とも呼ばれ、検索キーに関わるノードとエッジ（以下コアノード）をすべて削除する方法である。一般的にコアノードの共出現ウエイト（*cw*）の値がきわめて大きい時、すべてのノードはコアノードと結ばれ、放射線状に真っ黒な図形となる。これを自転車の車輪に喩えて、「スポークエフェクト」と呼んでいる。特に、地名のような特定の和歌にしか用いられない語の場合、よく見られる。そもそもコアノードにあたるキーワードで検索した歌のデータを用いてネットワークを描いているのであるから、すべての歌はコアノードと関係する。このことを前提に分析するなら、コアノードを刈り込んで見通し良くしてもかまわない。

図1は古今集のデータを用いて「梅」ネットワークを描いたものである。(a)は「梅」ノードの削除前、(b)は削除後である。古今集の場合、コアノードを削除しなくてもある程度、語相互のつながりは観察できるが、削除した方がよりわかりやすい。「梅」「鶯」「梅の香」「鶯が縫う梅の花笠」「梅花を折る」など、古今集特有の語のつながりが見えるようになった。

一方、図2は新古今集のデータを用いた「梅」ネットワークより「梅」ノードを削除する前(a)と削除した後(b)である。新古今集の場合、コアノードを削除しないと、「梅」以外の語のつながりは見えにくい。図2(a)をサッカーボールのような球体と見るならば、ボールの中心に「梅」があり、「梅」から伸びるエッジが球面を支えている（あるいは、つなぎとめている）ように見える⁶。「梅」を取り除くと、ちょうどボールの展開図が開くように、語相互のつながりが広がって見える(図2(b))。

⁶しばしば、図2(a)のような二重輪の構造（外側の輪と中心に集まるモコモコとした雲）になる。樹形図の描かれ方と同じなのであるが、その理由はまだよくわからない。

古今集の図1(b)と新古今集の図2(b)を比較すると、前者では「鶯」と「花」「色香」の関係が見えるのに対し、後者では、前者には見られなかった「鶯」と「雪」の関係が見える⁷。

3.2 2語の共有ノードの違い

つぎに、2語のネットワークの統合と交差による分析を示す。

図3(a)は「鶯」と「桜」の統合と交差、図3(b)は「鶯」と「梅」の統合と交差を示したものである。交差部分はグレーで示されている。

「鶯」は『万葉集』から数多く詠まれ、梅の花に鳴く鶯が最も多く、初春に鳴く鶯が春の最初に咲く梅の花とともに詠まれるのは当然といわれている[2, pp. 71-2]。図3が示すように、(a)と(b)の交差部分を比較すると、共に「鶯」と「桜」の各ノードはグレーで示されておらず、互いに同じ歌では出現しないことがわかる。共有するノードの数も3と少ない。一方、「鶯」と「梅」のノードは共にグレーで示されており、同じ歌に2語が使われていることがわかる。共有するノードの数も15であり、この2語の関係がよく詠まれることがわかる。

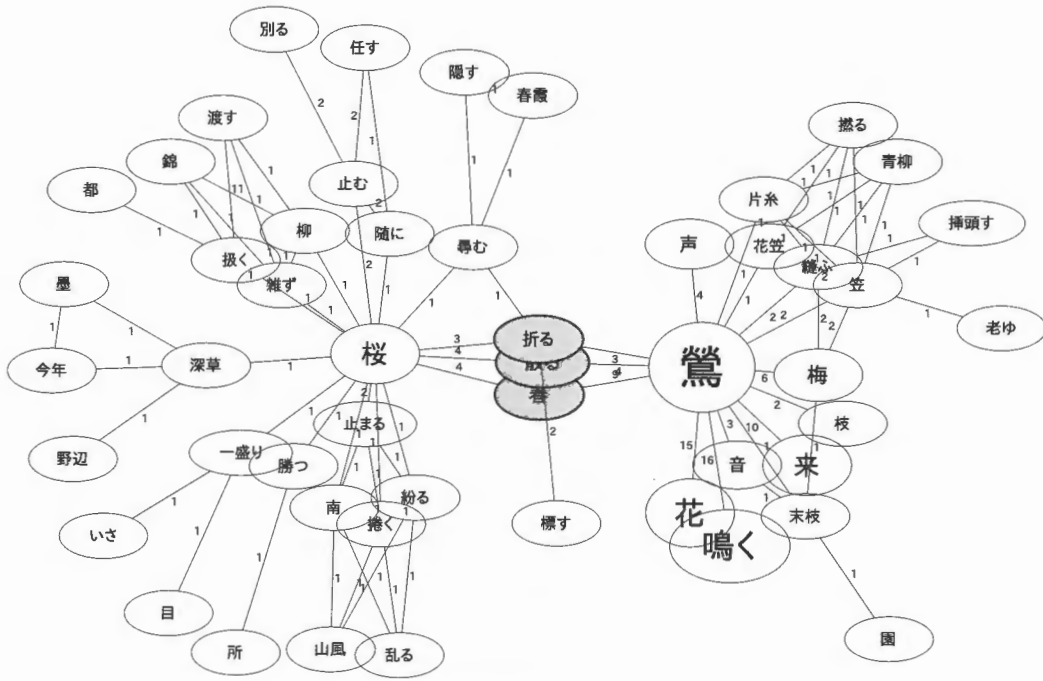
3.3 歌集で変化する共有ノード

最後に「桜」と「吉野」の関係が歌集によって変化することを示す。

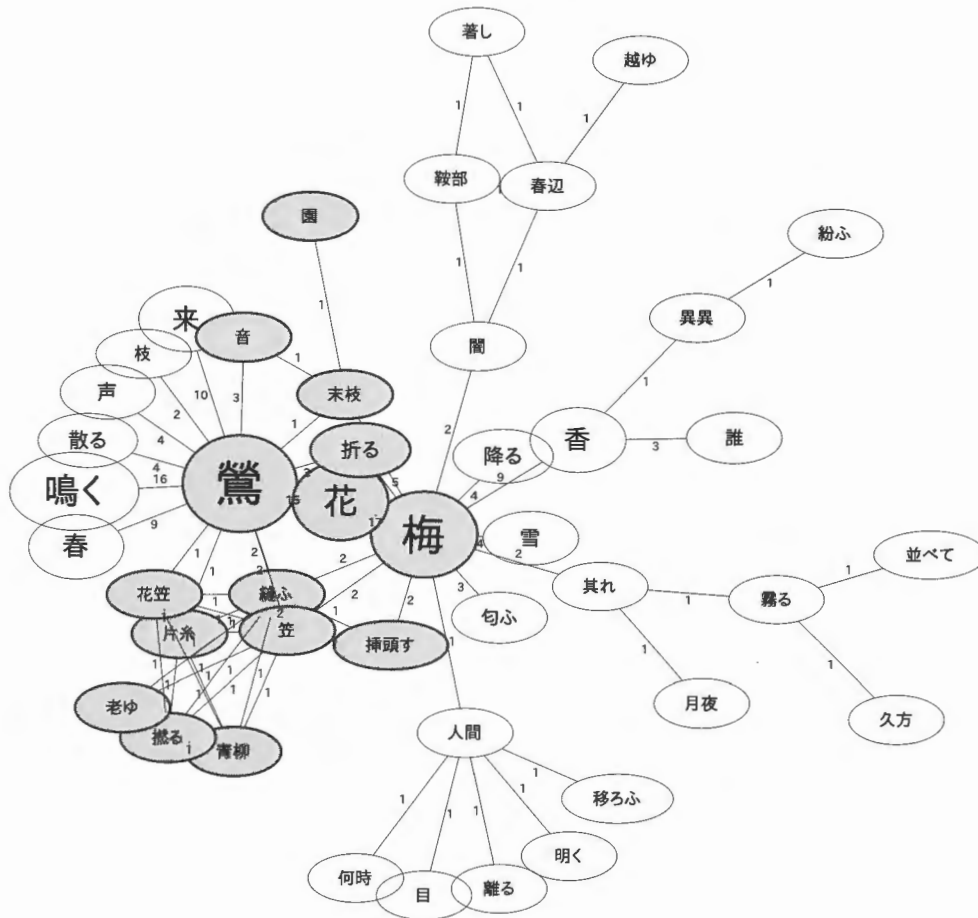
図4は、古今集における「桜」と「吉野」の関係(a)と新古今集における「桜」と「吉野」の関係(b)を示したものである。今でこそ「桜」と「吉野」の関係は有名であるが、「吉野山と桜の関係が決定的なものになるのは、やはり『吉野山去年こぞのしをりの道かへてまだ見ぬ方の花をたづねむ』(新古今集・春上)を代表とする数々の歌をよんだ西行とその時代[2, p. 436]で、古今集の時代では、「桜」と「吉野」の関係より「雪」と「吉野」の関係の方が強いといわれる[2, p. 435]⁸。図4(a)を見ると、確かに古今集の「吉野」は「桜」との関係よりも「雪」

⁷この関係は、新古今集30番（読人不知）「梅か枝に／なきてうつるふ／鶯の／はね白妙に／あは雪そふる」に見られる。

⁸片桐[2, p. 435]はよると「山岳信仰と結びついた吉野の山々のたたずまいがますます神秘的イメージになって行ったのであろう。(略)山岳信仰の地・隠遁の地としての吉野山であったが、そのような神秘的なイメージは雪をいただく山々の姿とマッチして、吉野山といえば雪がよまれるというようになった」という。



(a)



(b)

図 3: 古今集データにおける「鶯/桜」(a)と「鶯/梅」(b)の統合と交差

との関係の方が強く、「吉野」のネットワーク中に「雪」「白雪」「御雪」「寒し」のように「雪」を表す語や、「隠れ家」「(雪道) 踏み／平らす (馴らす)」のように「隠遁」を表す語が見られる。図4 (a) と (b) の2つの歌集 (約 905 年と 1205 年の成立) を比較することによって、「桜」と「吉野」の関係が時代につれて変化していることがわかる。

4 おわりに

本稿は、ブーリアン演算で語彙の集合を分析する方法について述べた。任意の2語の共出現パターンの違いを統合・交差を用いて示すことができた。また、同様の方法により、時代にわたって2語の関係の変化を示すことができた。どの演算を利用するかは、あらかじめ部分的に出力された図を見た上で、研究目的に応じて、適宜判断しなければならない。どの演算がどのような局面に有効であるかは、今後の課題としたい。

参考文献

- [1] 犬飼隆：平安末期複合動詞の意味構造, 国語語彙史研究会 (編), 国語語彙史の研究, 第9巻, pp. 272-258, 和泉書院 (1988).
- [2] 片桐洋一：歌枕歌ことば辞典, 角川小辞典, 第35巻, 角川書店, 東京 (1983).
- [3] 計量国語学会 (編): 計量国語学事典, 朝倉書店 (2009).
- [4] 国立国語研究所 (編): 分類語彙表/フロッピー版, 国立国語研究所言語処理データ集, 第5巻, 大日本図書, 東京 (1994), 『分類語彙表』は1964年に国立国語研究所資料集6林大担当として刊行された。
- [5] 近藤みゆき: n グラム統計処理を用いた文字列分析による日本古典文学の研究—『古今和歌集』の「ことば」の型と性差—, 千葉大学「人文研究」, Vol. 29, pp. 187-238 (2000).
- [6] 近藤みゆき: n-gram 統計による語形の抽出と複合語—平安時代語の分析から—, 日本語学, Vol. 20, pp. 79-89 (2001).
- [7] Manning, C. D. and Schütze, H.: *Foundation of statistical natural language processing*, The MIT press, Cambridge, Massachusetts (1999).
- [8] 水谷静夫: 共出現関係に拠る語彙分類の試み, 計量国語学, Vol. 77, pp. 1-13 (1976).
- [9] 水谷静夫: 語の共出現に拠る語彙構造探究の諸法, 計量国語学, Vol. 79, pp. 1-18 (1976).
- [10] 水谷静夫: 用語による梅・桜の歌の弁別, 計量国語学, Vol. 12, pp. 1-13 (1979).
- [11] 水谷静夫: 語彙, 朝倉日本語新講座, 第2巻, 朝倉書店, 第1版 (1983).
- [12] 中野洋: 新聞語彙調査の類別語彙表について, 電子計算機による国語研究 II, 国立国語研究所報告, 第34巻, pp. 38-54, 秀英出版, 東京 (1969).
- [13] 西端幸雄: 「歌物語」3作品の使用語彙の比較, 「歌物語」語彙の数量的分析と研究, pp. 3-18, 第1版 (1996), 文部省科学研究費: 重点領域研究「人文科学とコンピュータ」研究成果報告書.
- [14] Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF, *Journal of Documentation*, Vol. 60, pp. 503-520 (2004).
- [15] Rocchio, J. J.: The SMART Retrieval System: Experiments in Automatic Document Processing, in Salton, T. G. ed., *Relevance feedback in information retrieval*, pp. 313-323, Prentice-Hall, Englewood Cliff, NJ, 1 edition (1971).
- [16] 山田進: 意味分類辞書, 国語学, Vol. 53, No. 1, pp. 30-43 (2002).
- [17] 山元啓史: 古今集データベースによる歌語の視覚化, 人文科学とデータベース, 第11回シンポジウム, pp. 81-8, 人文科学とデータベース協議会, 大阪 (2005).
- [18] 山元啓史: コンピュータによる歌枕の分析, イタリア日本語教育協会, 第3回シンポジウム論文集, pp. 373-382, イタリア日本語・日本語教育学会 (2006).
- [19] 山元啓史: 歌ことばの可視化とコノテーションの抽出—グラフによる共出現パターンの作り方—, じんもんこん 2006, 人文科学とコンピュータシンポジウム, Vol. 2006, No. 17, pp. 21-28 (2006).
- [20] 山元啓史: ネットワークによる歌ことばのモデリング, 語彙研究, Vol. 2007, No. 5, pp. 21-32 (2007).
- [21] 山元啓史: モデリングによる歌ことばの変遷と分析—八代集・歌ことばシソーラスの開発—, じんもんこん 2007, 人文科学とコンピュータシンポジウム, Vol. 2007, No. 15, pp. 163-170 (2007).
- [22] 山元啓史: 和歌のための品詞タグづけシステム, 日本語の研究, Vol. 3, No. 3, pp. 33-39 (2007).
- [23] 山内洋一郎: 連歌分類語彙表 (体の類) 試案—宗祇関係千句連歌七種による—, 国語語彙史研究会 (編), 国語語彙史の研究, 第6巻, pp. 358-348, 和泉書院 (1985).
- [24] 田中章夫: 国語語彙論, 明治書院 (1978).