（研究ノート）

クラスター分析を用いた古典籍の分類

―源氏物語短編写本の調査―

齊　藤　鉄　也

キーワード

計量文献学　　源氏物語　　仮名字母　　頻度分布　　階層的クラスター分析

１．はじめに

計算機の性能の向上と、インターネット上の文書公開の普及によって、大量の文章を処理する統計的手法を用いた研究が増加している。本研究では、このアプローチを古典籍に適用し、その可能性を調査すること目的としている。具体的には、前稿[1]に引き続き、源氏物語の短編の写本115本を調査対象として、鎌倉時代初期から江戸時代初期の写本の仮名字母に着目し、階層的クラスター分析を用いて、写本の持つ傾向を明らかにすることである。古典籍を対象とする国文学分野に統計的手法を導入することで、国文学分野の中で定性的な議論の結果として生まれた仮説を検証し、その蓋然性を高めることができる可能性がある。

以下、第２章では、対象とする源氏物語の短編の写本の概要とその分析手法について述べる。第３章では、調査結果とその考察について述べる。第４章は今後の課題、第５章はまとめである。

２．調査概要

ここでは、対象とした源氏物語の短編写本の基礎情報を述べる。本調査では、仮名字母に関する情報を得て、文書間の類似度を明らかにする。そのために、統計的手法である階層的クラスター分析を用いる。

２－１．調査目的と調査方法

本調査の目的は、古典籍へ階層的クラスター分析を適用し、その課題を調査することである。クラスター分析は、教師なし分類方法のひとつであり、対象とするデータを分類するために用いる。そのため、課題を発見する探索的調査に向いている。

本調査では、古典籍の文字情報、特に古典籍に使用されている変体仮名の字母（字源）の出現頻度率に着目し、それを特徴量として用い、調査方法として階層的クラスター分析に用いている。特徴量である仮名字母の出現頻度率からは、生成されたクラスターの性質を明らかにすることは難しいので、クラスターに属する古典籍の共通の特徴からクラスターの性質を考えることとする。

さいとう　てつや：淑徳大学　経営学部　准教授

階層的クラスター分析で行う処理は、大きく二段階に分かれている。第一段階として、対象とするデータ間の類似度を計算する。類似度の計算方法は複数提案されている。第二段階として、類似度に基づいてクラスターを構成する。構成方法も複数提案されている。第一段階と第二段階それぞれにおいて、より良い分類結果を出力するための方法の選択は、対象とするデータの特徴と、利用者の経験と主観に依存している。

そこで、本調査では、対象としたデータに対して、文書検索の際に採用される類似度である、コサイン類似度を用いて、クラスター分析を行う。この分析結果に基づいて、生成されたクラスターの特徴を検討する。次に、様々に提案されている類似度の計算方法うち、参考文献[3]に掲載されている代表的な類似度計算方法を取り上げ、その結果の評価を行う。これらの処理は統計処理及びグラフィックスのための実行環境R[4]を用いて行う。

2－2．調査対象データ

調査対象とした写本は、参考文献[2]と同じ、源氏物語の短編写本115本である。調査対象の詳細は参考文献[2]に掲載しているため、ここでは概要を述べる。

対象とした源氏物語の写本は、源氏物語のうち短編を中心に、第３帖「空蝉」、第８帖「花宴」、第11帖「花散里」、第16帖「関屋」、第27帖「篝火」、第38帖「鈴虫」の６帖である。これらの帖（巻）の文字数は次の通りである。第11帖「花散里」、第16帖「関屋」、第27帖「篝火」は、およそ1500文字から2000文字程度の本文を持つ写本である。第３帖「空蝉」、第８帖「花宴」は、およそ4500文字から5000文字程度の本文を持つ写本である。第38帖「鈴虫」は、およそ6000文字から6500文字程度の本文を持つ写本である。

対象とした写本の冊数は、それぞれ第３帖「空蝉」は18冊、第８帖「花宴」は19冊、第11帖「花散里」は20冊、第16帖「関屋」は20冊、第27帖「篝火」は19冊、第38帖「鈴虫」は19冊であり、６帖の写本の合計冊数は計115本である。

資料からの採字の方針は、前稿[1]に基づいている。方針の要点は、本行本文の仮名を採字すること、仮名の字母表に基づいて、採字した仮名字母を分類すること、のふたつである。この方針に沿って採字された仮名字母は、同一の仮名（字音）を表す字母ごとに、その出現頻度率を計算する。この出現頻度率を用いて書写資料間の類似度を比較する。今回対象とした異なる巻の写本においては、巻ごとに本文が異なるため、単純な仮名字母の出現頻度数の比較では、その類似度を比較することは困難であることから、この出現頻度率を採用している。

3．結果と考察

ここでは、２章で述べた資料に対して、仮名字母の出現頻度率の類似度に基づいた調査結果を述べる。

2

3－1．類似度に基づく分布

仮名字母の出現頻度率を利用して、写本間の類似度を計測した。写本間の類似度の計算方法としては、情報検索において用いられることが多いコサイン類似度を用いて、調査した[3]。二つの文章のコサイン類似度を用いて、より似た写本間の類似度の値が小さくなるようにコサイン距離（非類似度）[6]を求める。ここでは、距離の値がコサイン値であるため、コサイン値を角度（度数）に変換して比較している。このため、距離が近い写本間の値は０に近くなる。

対象とした写本間の距離のヒストグラムを図１に表す。距離のサンプルサイズは6555である。

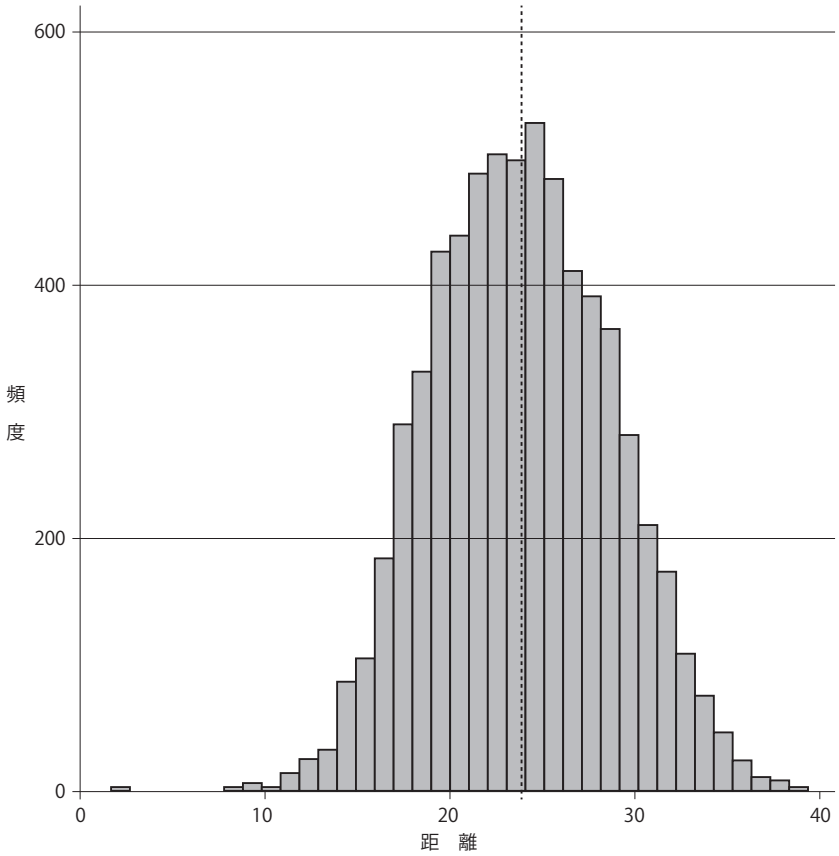

図１　源氏物語短編115写本間の距離のヒストグラム

距離の平均は23.86、標準偏差は4.85、距離の最小値は1.87、最大値は39.75であった。図１のほぼ中央にある点線が平均値を表している。

この分布に対して、コルモゴロフ－スミルノフ検定を用いて正規分布の検定を行ったところ、有意水準５％で棄却された。この結果、コサイン距離を用いた場合の類似度の分布は、正規分布に従っていないと言える。

図１からは、距離が10より近い組み合わせの数は極めて少ないことが明らかになった。このことから、ある一定の距離の値よりも近い写本の組み合わせは、仮名字母の出現頻度率が似ていることが想定される。ひとつの可能性として、それらの写本は「同一人物が書写した」または「異なる人物が元となる本（親本）を忠実に書写した」ことが考えられる。

3

３－２．クラスターの評価

さらに、写本間の類似度の傾向を確認するために、コサイン距離の値を用いて、階層的クラスター分析を行った。類似度の結合にはウォード法を用いた。この結果を図２に表す。

階層的クラスター分析では、分析結果として樹形図（デンドログラム）が描かれ、利用者の判断に基づいて適当な高さで樹形図を切断する。図２の場合であれば、高さ55から60の間で切断した場合、クラスターが三つまたは四つに分類することができる。ここでは、鎌倉時代、室町時代、江

戸時代の三つに分類することを想定し、三つに分割する。図２では点線によって切断した高さを表している。図２の資料名の最初の数字は、源氏物語の巻数を表している。03とは第３帖「空蝉」、08とは第８帖「花宴」、11とは第11帖「花散里」、16とは第16帖「関屋」、27とは第27帖「篝火」、38とは第38帖「鈴虫」を表している。巻数に続く文字列は所蔵者に基づく写本名を表している。

左側のクラスターには、鎌倉時代に書写された写本が多く集まっている。左寄り中央のクラスターには、江戸時代に書写された東久邇宮家旧蔵本が集まっている。右側のクラスターには、鎌倉時代と室町時代に書写された写本が多く集まっている。

右側のクラスターにある右から５番目に位置する中院文庫本は、日本大学蔵三条西家本の写本と類似度が近い。日本大学蔵三条西家本「空蝉」と「鈴虫」は三条西公条が書写している。中院文庫本は日本大学蔵三条西家本を親本として書写していることが知られている。日本大学蔵三条西家本「空蝉」と中院文庫本「空蝉」の類似度が近いことから、中院文庫本「空蝉」は日本大学蔵三条西家本「空蝉」を忠実に書写している、と考えられる。

この調査結果からは、仮名字母の出現率の視点から見て、鎌倉時代の写本は互いに類似している本があることが明らかになった。クラスター分析ではこれ以上の分析はできないが、今回の調査においては、仮名字母を分類のための特徴量としているため、鎌倉時代の写本には共通の仮名字母の使い方があることが推定できる。

３－３．誤分類率に基づく類似度の計算方法の比較

上記では、類似度の計算方法としてコサイン類似度を用いていた。これまでに文書間の類似度の計算方法は多く提案されている。そこで、ここではRの類似度計算パッケージの持つ類似度のうち、参考文献[3]に掲載されている代表的な三つの類似度計算方法、コサイン類似度、相関係数、拡張ジャッカール係数を取り上げ比較する。

比較方法は、次の手順で行う。最初に、上記と同様に類似度を計算する。類似度の計算方法は代表的な三つの類似度計算方法を用いる。次に、ウォード法に基づいてクラスターを構成する。クラスターは３－２と同様に構成されたのでクラスターを三つに分割する。この際に、調査対象とした写本の推定書写年代とクラスターの分類結果とのクロス集計を行い、この集計結果を評価する。評価では、写本の持つ鎌倉時代、室町時代、江戸時代といった書写年代と、主に鎌倉時代の写本が集まるクラスター、主に室町時代の写本が集まるクラスター、主に江戸時代の写本が集まるクラスターのクロス集計の結果、書写された時代と時代を表していると考えられるクラスターが合致する場合を正しく分類された、と考える。この正しく分類された結果を用いて、誤分類率は計算し、その分類結果を評価する。誤分類率は低い方が望ましい。

誤分類率の計算結果を表１にまとめた。表１では、三つに分けたクラスターが縦方向に並んでいる。クラスター１に主として鎌倉時代の写本が集まっている。クラスター２に主として室町時代の写本が集まっている。クラスター３に主として江戸時代の写本が集まっている。横方向にそれぞれの類似度の計算方法名と、写本の書写年代を並べている。例えば、コサイン類似度の場合、鎌倉時代に書写されたとされる写本、計28本のうち、主として鎌倉時代の写本が分類されているクラスターに分類された写本は23本あることを表している。書写年代とクラスターが合致し、正しく分類された写本数は右上から左下に向けて対角線方向に並んでいる。この正しく並んでいる写本数を用いて、一番下の段の誤分類率を計算している。

比較した結果は、コサイン類似度、拡張ジャッカール係数、相関係数の順に誤分類率が低い結果となった。推定書写年代には、幅のある年代が記述されている写本も多い。例えば、鎌倉時代後期

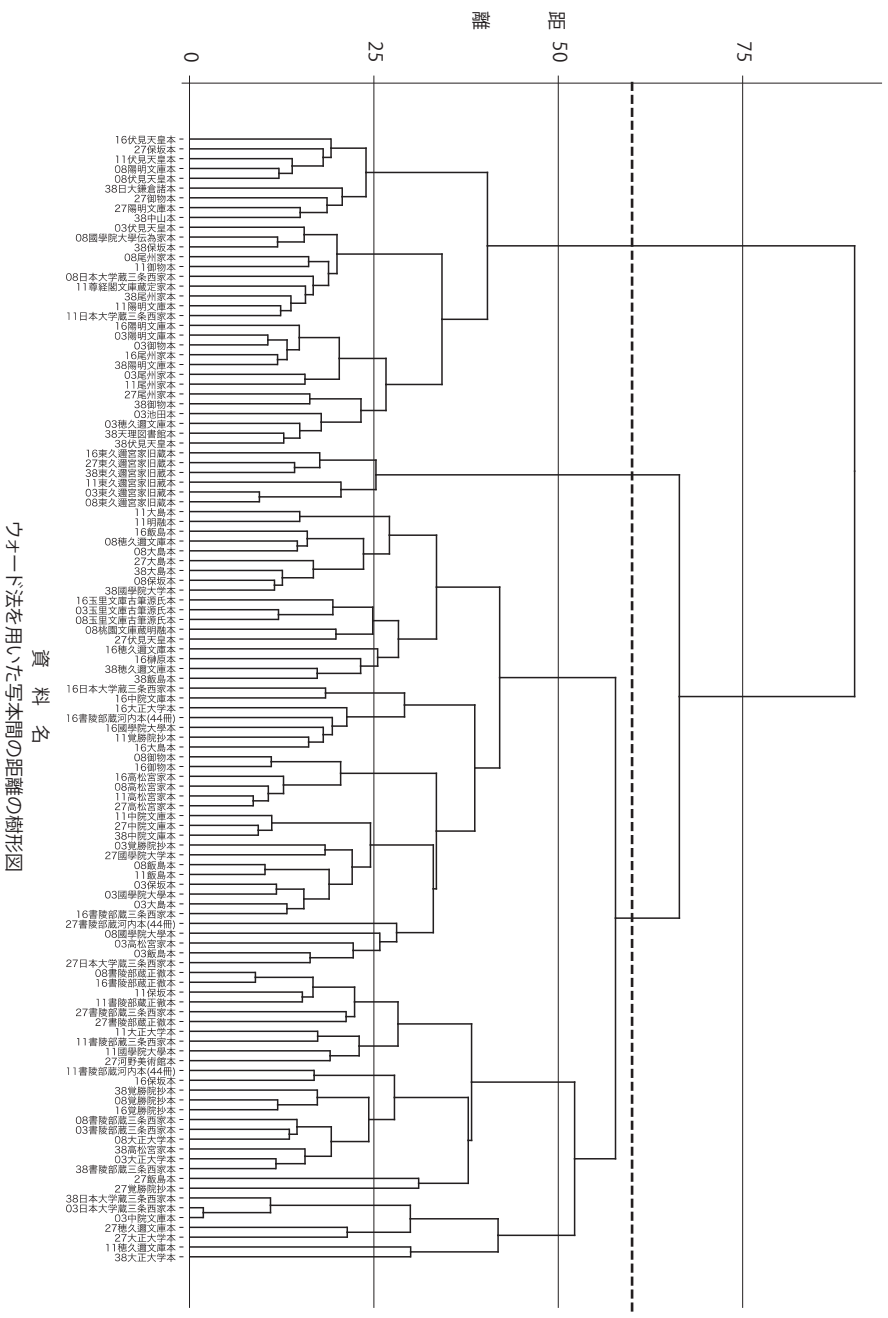

図２　源氏物語短編115写本の階層的クラスター分析による樹形図

資料名
ウォード法を用いた写本間の距離の樹形図

表１　類似度の計算方法の誤分類率

クラスター	コサイン類似度			相関係数			拡張ジャッカール係数		
	鎌倉	室町	江戸	鎌倉	室町	江戸	鎌倉	室町	江戸
1	23	9	0	24	17	0	24	11	0
2	5	66	6	4	58	6	4	64	6
3	0	0	6	0	0	6	0	0	6
誤分類率	1−(23+66+6)／155=0.174			1−(24+58+6)／115=0.235			1−(24+64+6)／115=0.183		

または南北朝時代、といった記述もある。この場合には、より後の時代の分類している。加えて、南北朝時代は室町時代前期と重なっているため、本調査においては、室町時代前期と分類している。そのため、結果として誤分類として判断された写本も一定数含まれる。また、室町時代に書写された桃園文庫蔵明融本「花宴」は、鎌倉時代に書写された定家筆本を忠実に書写した本とされている。そのため、書写年代は室町時代後期ではあるが、この場合、鎌倉時代に書写されたと見做して良いと考えられる。これも誤分類として判断されている。これらのことから、実際の誤分類率はより減少すると考えられる。江戸時代に分類されている写本は、全て東久邇宮家旧蔵本であることから、どの類似度計算方法においても区別が可能な特徴的な仮名字母の出現頻度率を持つことが想定される。

文章間の類似度の計算にはコサイン類似度が多く用いられている。この結果も、古典籍においても、コサイン類似度の誤分類率が低く、その有効性を示したと言える。上記にも述べたように、類似度の計算方法はこの他にも多数存在するので、今後、様々な類似度の計算方法を用いて、より精度の高い計算方法を調査する必要がある。

４．今後の課題

クラスター分析の利点として、その結果が樹形図（デンドログラム）として可視化され、類似度に基づいてデータ間の関係が明らかになることから、利用者にとって理解し易いことがある。但し、生成された樹形図は計算結果によって変化するため、客観的な根拠として用いることはできないこと、その評価は利用者の主観に依存すること、といった制限がある。その資料の情報に基づいた内的な根拠による計算結果に加えて、樹形図から仮説を生成し、その後、別の外的な根拠に基づいて検証することが必要となる。

古典籍に適用する場合において、その書写者は、書写者が不明もしくは古筆鑑定に基づく伝承筆者として名前が伝わる書写者の情報しかないことが多い。その点で、外的な根拠に基づいて検証されたクラスターに対して、情報が限られている古典籍を分類することは、ひとつの方法として有効である可能性がある。

これに対し、クラスター分析の欠点は、マクロな視点に基づいて類似度を比較しているため、その詳細が不明なことである。例えば、今回の調査の場合、ある写本間の類似度が近いことが明らかになったとするならば、類似度が高い写本の仮名字母の調査をより詳細に行うことで、それらの写本に関する特徴が明らかになると考えられる。残念ながら、クラスター分析では、類似度の計算結果に基づいているため、このよりミクロな視点に基づいて分析することは不可能であり、異なる手法を用いて分析する必要がある。そのため、クラスター分析により、何らかの仮説を提案することは可能であるが、その根拠を明らかにするためには別の手法による分析が必要である。

5．まとめ

本研究では、統計的処理を古典籍に適用して、仮名字母の類似度の視点から、源氏物語の短編写本を調査した。これまでの情報学または統計学の分野の研究者が古典籍を対象とした研究は、その分野の目的に基づき、数理的な方法の適用とその結果の評価分析に重きを置いた。そのため、調査対象としたデータは現代の通行仮名に翻刻された文章を用いているため、古典籍を研究する研究者との接点が狭くなる傾向があった。一方、国文学の分野の研究者が古典籍を対象とした研究は、単純な数値の比較や、これまでの知見や経験に基づいた定性的な議論が行われ、数理的な処理に基づく定量的な結果との議論が困難であった。この点で、学際的なアプローチに基づいた研究を行うことが難しかった。

これに対し、本研究では、原典である写本から変体仮名を直接翻字した仮名字母の情報を用い、統計的処理を用いて分析している点で、これまでの研究とは異なる方法を採用している。その結果、対象とした写本の中において、仮名字母の出現頻度率の類似度と書写年代の関係に関して新しい知見を得たと言える。

本調査では、階層的クラスター分析を用いて、源氏物語の短編写本間の類似度を明らかにし、鎌倉時代に書写された写本の中に類似度の高い写本群があることを明らかにした。今後は、多く残され、かつ、出版やインターネット上に公開されている写本を対象に、今回と同様の手法の適用や手法の検討を行っていきたい。

謝辞

調査と研究を進める際に様々な助言と示唆を与えていただいた、尾州家源氏物語研究会と早稲田大学文学学術院陣野研究室の皆様に感謝いたします。

参考文献

［1］齊藤鉄也、「源氏物語『篝火』本文の文字の分布」、国際コミュニケーション学会　国際経営・文化研究、Vol.20, No1, 2015

［2］齊藤鉄也、「計量文献学の古典籍への応用―データマイニング手法を用いた源氏物語短編写本の分析―」、投稿中、2016

［3］Alexander Strehl, “Similarity Measures for Document Clustering”, http://www.strehl.com/diss/node52.html, 2002

［4］The R Foundation, “The R Project for Statistical Computing”, https://www.r-project.org/

［5］村上征勝、「真贋の科学―計量文献学入門」、朝倉書店、1994

［6］金明哲、「テキストデータの統計科学入門」、岩波書店、2011

（受理　平成28年９月６日）