1G-4

校本に基づく諸本の階層的分類と伝本系譜の推定

野村 厚志† 熊本 守雄‡
山口大学教育学部† 山口県立大学名誉教授‡

1 はじめに

諸伝本の異同を調べる校合の作業を通じて、それらの系譜を推定することが、研究者らによって行われる。この推定は、研究者の知識・経験や直感に基づいて進められる。伝本系譜の推定が計算機で可能となれば、定量的な裏付けが得られることや、人間が見落としているかもしれない知見・発見が得られることが期待できる。

ここでは、既存の校合の結果をまとめた校本を用いて、諸本の異同のデータから、諸本の類似度を計算し、類似度データを基に階層的[1]に似通った諸本を分類することを試みる。その上で、分類結果に、全文字数に対する漢字の割合を表す漢字率を指標として加えることで、伝本系譜を推定する情報処理手法を提案する。提案した手法を「遍昭集」の校本データ[2]に適用し、その結果を紹介する。

2 諸本間の類似度評価

今、N 個の諸本があり、それらに番号 i を付加する(i=1,2,...,N)。既にそれらの諸本間の校合結果があり、M 個の句に分割されて異同が調べられている仮定する。これらの句に番号 m=1,2,...,M を付加する。また、諸本 i の句 m を $X_{i,m}$ とあらわすことにする。

まず、2 つの諸本 i と j との句 m に関する差異を次のように定義する。

$$\delta(X_{i,m},X_{j,m})=\begin{cases}1 & \text{if} \quad X_{i,m}=X_{j,m}\\ 0 & \text{otherwise}\end{cases} \qquad (1)$$

すなわち、差異がなければ 1、何らかの差異があれば 0 となる。式(1)は 1 つの句についての差異を表すので、すべての句に対して平均することにより、2 つの諸本間の類似度 $C_{i,j}$ を評価する。

$$C_{i,j}=\frac{1}{M}\sum_{m=1}^{M}\delta(X_{i,m},X_{j,m}) \qquad (2)$$

ここで、$C_{i,j}=C_{j,i}$, $C_{i,i}=1.0$, $0\le C_{i,j}\le 1$ である。

3 階層的分類

まず、N 個の諸本を、式(1)及び(2)を用いて計算された類似度に基づいて分類する。伝写の過程において、意図しない誤りや、仮名書を漢字に改めるなどの改変が含まれることがある。類似度の大きい 2 つの諸本は似通っており、小さいものは大きく異なっている。従って、類似度の相対的な大小は、諸本の中での親子関係を示唆するものと期待できる。そこで、類似度に従って諸本を階層的に分類していくことにより、伝本系譜を推定するための分類をおこなう。階層的分類（クラスタリング）のためのいくつかのアルゴリズムが提案されているが[1]、ここでは、最短距離法に基づく分類法を採用する。

具体的な階層的分類の処理方法を以下に示す。

まず、N 個の諸本を N 個のクラスタとする。N 個のクラスタ間で最も類似度の大きいもの 2 つを探し出し、それらを融合して新たな 1 つのクラスタとする。この時点で、クラスタ数は(N-1)個となる。次にその(N-1)個のクラスタ間で、再度、類似度を計算する。その類似度から、最も類似度の大きい 2 つを探し、それらを 1 つのクラスタとする。このクラスタリングの処理を、クラスタ数が 1 となるまで再帰的に繰り返す。

クラスタ間の類似度の評価は、最短距離法を用いる。すなわち、2 つのクラスタそれぞれに含まれる任意の諸本の組み合わせについて、その類似度の最大値をそれら 2 つのクラスタの類似度とする。例えば、あるクラスタαに諸本 i と j が含まれ、別のクラスタβには諸本の k のみであったとする。このとき、2 つのクラスタαとβとの間の類似度 $C_{\alpha,\beta}$ は、次のようになる。

$$C_{\alpha,\beta}=\max\{C_{i,k},C_{j,k}\} \qquad (2)$$

最後に、階層的クラスタリングの結果から、伝本の系譜を推定する。伝写の過程において、仮名が漢字に置き換えられることがあり、全文字数に対する漢字の比率：漢字率が、親子関係の判定に有用な情報を与えると考える。そこで、親子関係にあると思われる 2 つの諸本間で、漢字率の低い方を親本と推定する。また、全ての諸本について、漢字率の最小のものを底本と推定し、伝本の系譜を推定していく。

Hierarchical clustering of manuscripts with an annotated textbook and estimation of their pedigree-chart
† Atsushi NOMURA (Faculty of Education, Yamaguchi University)
‡ Morio KUMAMOTO (Professor Emeritus, Yamaguchi Prefectural University)

4　伝本系譜の推定例

ここでは「遍昭集」を例として、その校本から伝本系譜を推定するまでの過程を紹介する。

遍昭集のうち、22 の諸本が現存しており、その校本が文献[2]である。具体的には、西本願寺本(i=1 と番号を付与)を底本として、他の諸本との比較によって、異同が記されている。例えば、冒頭の歌は、西本願寺本では次のように記されている（但し、句の区切りを「／」で記した）。

春／
はなのいろは／かすみにこめて／みせすとも／
かをたにぬすめ／はるのやまかせ／

これに対して、他の 21 の諸本でどのように記されているか、その異同が文献[2]において次のように記されている。

春 1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,20,21,22－ナシ 15／はなのいろは 1,2－はなのいろハ 6,22－花の色は 8,19,20－花の色ハ 3,4,5,7,9,10,11,12,13,14,16,17,18,21－花の色(香)ハ 15／かすみにこめて 1,2,5－かすミにこめて 6,7,13,22－霞にこめて 3,4,8,9,10,11,14,15,16,17,18,19,20,21－霞に籠て 12／みせすとも 1,2,3,4,5,10－ミせすとも 6,14－見せすとも 8,22－みえ(せ、)ずとも 11－見せ(え)すとも 12－みえすとも 7,9,16,17－見えすとも 13,18,19,20,21－見えす共 15／かをたにぬすめ－1,2,3,5,6,7,8,9,10,12,14,15,16,17,18,19,20,21,22－かをだにぬすめ 11－香をたにぬすめ 4,13／はるのやまかせ 1,2－春のやまかせ 6－春の山かせ 4,15,18,19－春の山風 3,5,7,8,9,10,11,12,13,14,16,17,20,21－春山風 22／

ここで、句の後の整数は諸本の番号 i を表し、また、ナシはその句がないことを表す。

上記で示した校合のデータについて、本研究で提案する推定方法に当てはめると、N=22、M=6 であり、例えばその中の底本 i=1 と諸本 i=3 の最初の 2 句 (m=1,2)は次のようになる。

$X_{1,1}$="春"　　$X_{1,2}$="はなのいろは"
$X_{3,1}$="春"　　$X_{3,2}$="花の色ハ"

また、これらの句の差異は次のようになる。

$\delta(X_{1,1}, X_{3,1}) = 1,\ \delta(X_{1,2}, X_{3,2}) = 0$

実際に、M=199 個の校合データについて、類似度を計算した結果を図 1(a)に、さらに、階層的に諸本を分類した結果を図 1(b)に示す。諸本の 1 と 2、19 と 20 は極めてよく一致しており、それぞれ一つのクラスタとして分類されていることが図 1(b)からも確認できる。

さらに、各諸本において漢字率を計算し[図 1(c)]、上の結果と合わせて伝本系譜を推定した最終的な系譜図が図 1(d)である。校本では諸本の 1 が底本として扱われており、諸本の 2 はそれを忠実に転写した模写本と考えられている。提案したアルゴリズムにおいても、諸本の 1 が諸本の 2 の親本であり、なおかつ、底本であることが推定されている。

5　まとめ

ここでは、既存の校本に基づいて、諸本の類似度を計算し、階層的に分類すること、及び漢字率をも付加することで、伝本系譜を推定する情報処理手法を提案した。さらに、提案手法を「遍昭集」に対して適用した結果を紹介した。今後の課題として、まずは、類似度の評価方法を検討したい。式(1)のように校本の情報を十分利用しておらず、検討の余地がある。

参考文献

[1] 神嶌敏弘："データマイニング分野のクラスタリング手法(1)", 人工知能学会誌, **Vol. 18**, pp. 59-65, 2003

[2] 熊本守雄："校本・遍昭集", 山口女子大学國際文化學部紀要, **Vol. 1**, pp. 15-47, 1995

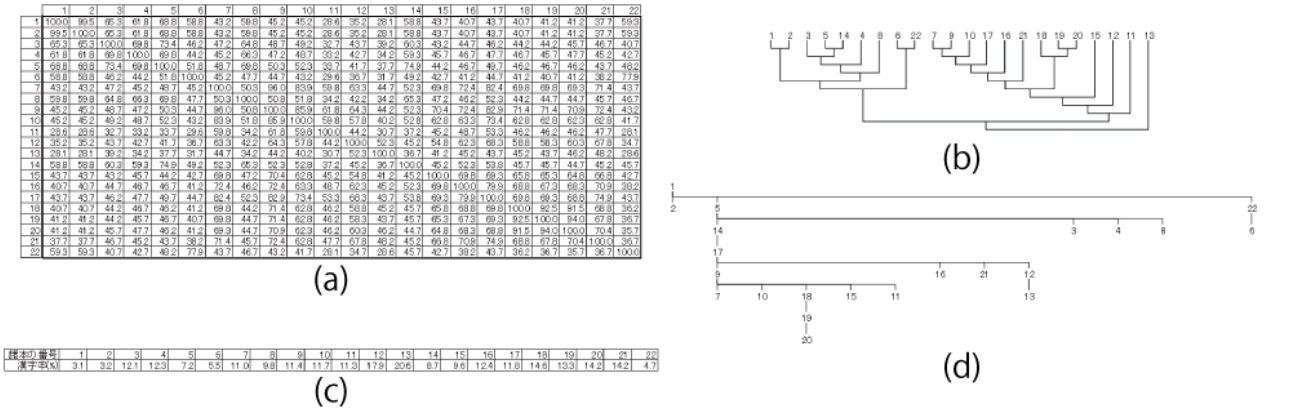

図１．「遍昭集」の校本[2]に基づく諸本の階層的分類と伝本系譜の推定結果。
(a)諸本間の類似度、(b)階層的分類結果、(c)漢字率、(d)伝本系譜の推定結果。