助詞分布における書き手の識別ルールの抽出

金 明哲(Jin Mingzhe)
jin@earth.sgu.ac.jp
札幌学院大学

1. はじめに

文章(文書)から書き手の文体の計量的な特徴を抽出し，その情報に基づいて文章の書き手を推定する研究では，欧米文では文の長さ，単語の長さ，単語の使用頻度などに関する情報がよく用いられている(Holmes, D. I. 1994)。日本文に関しては，文の長さの分布，品詞の使用率等が書き手の特徴情報としてよく用いられている(村上 1994, 安本 1994)。これらは，書き手の特徴を表す情報の一つであることは否定しないが，書き手によっては特徴が明確に現れない場合がしばしばあることが実証されている(金 1994a)。

文章のどのような要素に書き手の特徴が明確に現われるかに関しては，言語の種類によって異なる。そこで，筆者は日本現代文における書き手の特徴情報の抽出に関する研究に取り組み，いくつかの研究成果をあげている(Jin, M. and Murakami, M. 1993, 金 1993a～1997)。

本稿では，その一環として3人(井上靖，三島由紀夫，中島敦)の合計 28 の文章(126844単語)を用いて，品詞のなかで，使用頻度が最も高い助詞(約30%－40%)に注目し，

(1)助詞分布には書き手の特徴が明確に現われるか

(2)知識発見法に基づいて書き手の識別ルールの生成が可能か

について行った実証研究の結果を述べる。

2. 助詞の分布と書き手の特徴

本研究では出現頻度が高い助詞 23 種類とその他の助詞は 1 つのカテゴリーにまとめ，合計 24 項目(か，が，て，で，と，に，の，は，ば，へ，も，や，を，から，だけ，ても，でも，とも，ので，ほど，まで，ながら，ばかり，その他)に分けた，相対出現率を用いて分析を行った。統計数理のアプローチで様々な角度から分析を行った結果，助詞の分布には書き手の特徴が見られることが明らかにされた(金 1996a, 1996b, 1997)。

文章の中から抽出したデータを用いて書き手を推定・識別するためには，その情報を用いて文章を分類する場合，文章が書き手毎に分類されることが望ましい。分類分析にはいくつかの方法があるが，本研究では分散共分散の行列に基づいた主成分分析方法を用いて文章の分類を行った。その結果，第 1 主成分，第 2 主成分，第 3 主成分の寄与率はそれぞれ 31.15%, 29.29%, 10.21%で，第 3 主成分までの累積寄与率は 70.65%である。

図 1 に第 3 主成分までの 3 次元の主成分

得点の散布図を示した。主成分得点の３次元散布図では，文章が書き手別にはっきり分類されることがわかる。

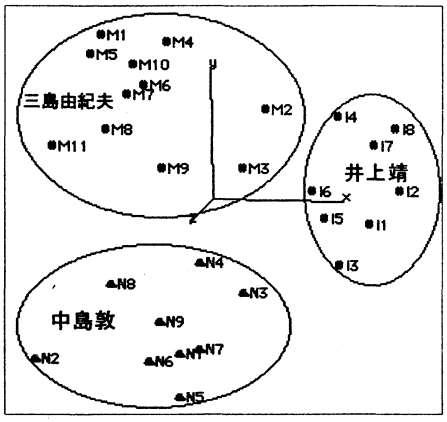

図１．助詞の分布に基づいた
文章の３次元散布図

主成分分析は学習データなしの分類方法である。学習データありの分類方法としては判別分析，ニューラルネットワークによる分類法等の方法があるが，判別分析がより広く知られている。判別分析にも多くの方法が提案されているが，本稿ではもっとも頑健な方法である距離法による判別分析を用いる(村上・金 1998)。表１にその結果を示す。表から分かるようにすべての文章が書き手毎に正しく判別されている。

表１．　助詞の分布を用いた判別結果

著者名	井上	三島	中島
井上	8	0	0
三島	0	11	0
中島	0	0	9

3. 書き手の識別ルール

助詞分布における書き手の特徴に関する分析は統計的のアプローチでは t, F, χ^2 等の統計量を用いて分析を行うことが考えられる。このような方法は分布の中のどの変数(項目)に書き手の特徴がより明確に現れるかにつて考察することが可能であるが，書き手を識別するルールを抽出するまでには至っていない。そこで本研究では近年注目を集めつつある知識発見法を用いて助詞分布における書き手を識別するルールの抽出を試みることにする。知識発見法にも様々なアプローチで研究が進められているが，ここではラフ集合理論に基づいた知識発見法を用いる。

ラフ集合(rough sets)の概念はポーランドの計算機科学者 Pawlak が 1982 年に提案された(Pawlak, Z. 1982) 。ラフ集合理論はファジィ集合とは異なり，従来のクリスプ集合と一線を引く定義はない。ラフ集合の概念はクリスプ集合上の「類別」と「近似」である。ラフ集合の定義を与えるため，全体集合 $U=\{S_1,S_2,\cdots S_{11},S_n\}$ 上に集合 X があるとする。いま S_i を用いて X を近似することを考える。S_i を用いた X の近似は

(1) X に完全に含まれている S_i を用いて X を近似

(2) X を含む S_i を用いて X を近似

二つの方法がある。方法(1)による X への近似は下近似，方法(2)による X への近似は上近似と呼び，それぞれ $\underline{A}X$, $\overline{A}X$ と記する。

X の下近似値と上近似値が一致しない場合，集合 X をラフ集合という。ラフ集合では，属性集合 A における下記のような下近似と上近似の関係式を

$$\alpha_A(X)=\frac{Card\ \underline{A}X}{Card\ \overline{A}X}$$

を用いて近似の度合を示す。式のなかの

$Card\ \underline{A}\ X$，$Card\ \overline{A}X$はそれぞれ下近似，上近似集合のなかに含まれた要素の数である。明らかに$\alpha(X)$は$0 \le \alpha(X) \le 1$であり，$\alpha(X)$が大きいほど近似の精度がよい。

Ziarkoはこのような概念に基づいて，分類を行なうときの判別精度の度合

$$E(B,C)=\begin{cases}1-\dfrac{Card(B \cap C)}{Card(B)} & if\ \ Card(B)>0\\ 0 & if\ \ Card(B)=0\end{cases}$$

を提案した(Ziarko, W., 1993)。$E(B,C)$の値をいき値というが，いき値が高いほど分類がよい。識別・判別のルールは選択された上位ランクの項目の組み合わせにより構成される。

本研究ではこのような理論に基づいて書き手毎の識別ルールを抽出するプログラムを作成し，井上靖，三島由紀夫，中島敦の28文章における助詞分布の全ての変数について分析を行った。

ラフ集合理論に基づいた知識発見のアルゴリズムでデータを解析するときには，まずいき値を決める必要がある。いき値は試行錯誤でデータと対話しながら決める必要がある。

上記のデータではいき値を0.85とすると，井上と三島，井上と中島が識別できる確率が0.85を超える助詞「か」，「て」，「と」，「に」が候補として選択される。最も少ない変数(助詞)，かつ最も高い識別率で井上と他の作家を識別するルールは下記のルール1，2である。

ルール1

if $(0.66 < x_1 < 2.86) \& (11.42 < x_3 < 15.21) \& (x_6 < 11.27)$
then 井上

ルール2

if $(11.42 < x_3) \& (6.19 < x_5) \& (x_6 < 11.27)$
then 井上

ここのx_1, x_3, x_5, x_6はそれぞれ「か」，「て」，「と」，「に」の使用率である。上記の二つのルールによる識別率はいずれも100%である。またここで抽出されたルールはいずれも井上と三島・中島を同時に識別可能なルールである。井上と中島だけを識別するなら$x_6 < 11.27$だけで十分である。同じな方法で求めた三島，中島を識別するルールの一部を下記する。

ルール3

if $(x_1 < 0.91) \& (3.63 < x_5 < 7.28) \& (0.71 < x_{19})$
then 三島

ルール4

if $(11.57 < x_6 < 14.41) \& (x_{19} < 0.20)$
then 中島

ルールの中のx_{19}は「ので」の使用率である。このようなルールは多く存在するがここで取り上げたのは最大の識別率でかつ最も少ない変数(項目)を用いたルールである。図2に井上を識別するルールに用いた3つの変数を用いた3次元散布図を示す。図で分かるように文章が書き手毎に分類されていることが分かる。

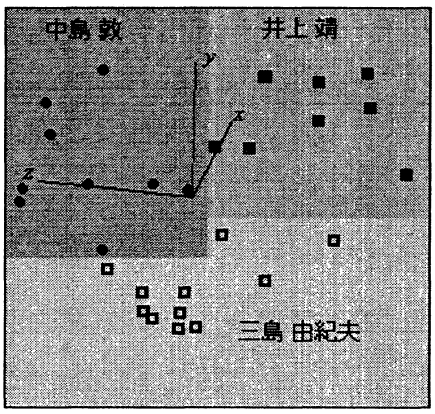

図2． 3つの助詞による3次元散布図

4. 終わりに

本研究では，品詞別に見た場合，文章の中で出現頻度が最も高い助詞に着眼し、助詞の分布に書き手の特徴が明確に現れるか否かを実証する同時に，知識発見法を用いて助詞の分布における書き手を識別するルールを生成する試みを行った。

その結果助詞の分布には書き手の特徴が明確に現れ，知識発見法により書き手を識別・判別するルールを生成することが可能であることが明らかにされた。ただし，大きくないサンプルサイズを用いて生成されたルールを用いて，書き手不明の文章の書き手を識別・判別することを考えた場合，その識別・判別率は判別分析より劣ることは十分予測できる。しかし生成された書き手の識別ルールは書き手の特徴をつかめるための非常に有益な情報であることは疑いない。

謝辞：知識発見法に関する研究にきっかけと情報を与えて下さった，関西学院大学の雄山真弓，岡田孝教授にお礼を申し上げます

参考文献

Holmes, D. I.(1994). Authorship Attribution, Computers and the Humanities, 28, 87-106.

Jin,M and Murakami,M(1993).Authors' Characteristic Writing Styles as Seen Through Their Use of Commas, Behaviormetrika, Vol.20, 63-76.

金明哲,樺島忠夫,村上征勝(1993a). 読点と書き手の個性,計量国語学,Vol.18, No.8, 382-391.

金明哲,樺島忠夫,村上征勝(1993b).手書きとワープロによる文章の計量分析,計量国語学,Vol.19, No.3, 133-145.

金明哲(1994a). 自然言語におけるパターンに関する計量的研究，総合大学院大学，学位論文.

金明哲(1994). 読点の打ち方と文章の分類,計量国語学, Vol. 19, No. 17, 317-383.

金明哲(1995).動詞の長さの分布に基づいた文章の分類と和語および合成語の比率,自然言語処理, Vol.2, No.1, 57-75.

金明哲(1996a). 動詞の長さの分布と文章の著者,社会情報，札幌学院大学社会情報学部紀要, Vol. 5, No.2, 13-22.

金明哲(1996b). 小説文における文節の係り受け距離の統計的特徴，計量国語学, Vol. 20, No. 4,168-179.

金明哲(1996c). 文節の係り受け距離の統計的分析，社会情報, Vol. 5, No.2, 1-12.

金明哲(1996d). 助詞分布に基づいた文章の著者の認識,
人文科学における数量的分析論文集(文部省科学研究費・重点領域研究), 49-54.
行動計量学会第 24 回大会論文抄録集,144-147.

金明哲(1997). 助詞の分布に基づいた日記の書き手の認識，計量国語学，20巻8号，357－367.

村上征勝(1994). 計量的文体研究の威力と成果,言語,Vol. 23,No. 2, 30-37 .

村上征勝・金明哲(1998). 「人文科学とコンピュータ」講座第5巻「数量的分析編」，尚学出版.

Pawlak,Z.(1982) : Rough Sets. International Journal of Computer and Information Sciences, 11, 341-356.

Pawlak, Z.(1984):Rough Classification, Int. J. Of Man-Machin Studies,20, 469-485.

安本美典(1994). 文体を決める三つの因子，言語, Vol. 23, No. 2, 22-29.

Ziarko, W.(1993). Variable Precision rough Sets Model. Journal of Computer and System Science, Vol.46, No.1, 29-59.