CD-003

多様な人物が執筆したテキストデータを対象としたその人物の性向を表す語彙・文体特徴の抽出方式

A Lexical and Stylistic Feature Extraction Method for Text Data Written by a Variety of People that Represents the Person's Propensities

小椋 佳歩† 岡田 龍太郎† 峰松 彩子† 中西 崇文†
Kaho Ogura† Ryotaro Okada† Ayako Minematsu† Takafumi Nakanishi†

1. はじめに

一般に，人物が執筆した文章には作者の経験や性質に起因した特徴が現れるとされている．これまで，作者推定に関する研究[1][2]がなされているという現状を鑑みても，文章に作者個人の性向が顕在化するということは明らかである．ここで，我々は図 1 に示すような執筆者経験考え方仮説を立てた．この仮説は文章に現れる文体及び使用語彙の特徴から最終的に作者の経験や思想を把握することができるというものである．作者の文体及び使用語彙を文章に現れる抽出可能な特徴とし，各特徴量から作者の文章の傾向という伝え方タイプを判別する．割り当てられた伝え方タイプより作者の文章を構成する筋道のことを指す文章構成方策を導出する．導出された文章構成方策から作者と相手との関係性及びそれを踏まえ何を伝えたいのかという関係伝達機能を経て，作者の経験や考え方を把握することが仮説の立証であり一連の研究の目的である．ここで，本仮説において文体とは文章の内容に依らない文章を構成する各要素に着目した数値によって判別可能な文章の表現方法であると定義する．また，使用語彙は文章内における各語彙の扱われ方であると定義する．

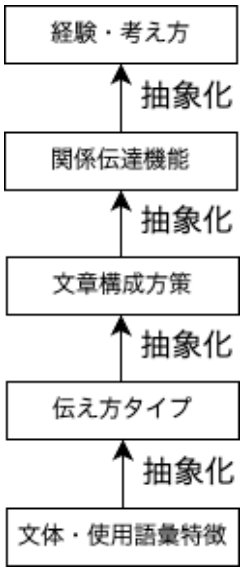

図 1　執筆者経験考え方抽出仮説

本稿では先に述べた仮説の立証へ向けた第一段階として作者の文章における文体及び使用語彙の特徴と伝え方タイプを可視化することを目的とした方式の提案をする．作者の文体及び単語特徴を抽出した上で，我々が提案する Approximate Inverse Model Explanations (AIME)[3]を用いて文章から作者を決定付ける際の各特徴の重要度を導出する．導出された結果をクラスタリングすることで作者の文章の伝え方タイプの可視化を実現，提案する．本方式によって作者が文章を書くうえでどのような特徴をもっており，どのような伝え方を選んでいるのかを客観的に把握することが可能になる．

2. 関連研究

Zhao[1]らは語彙的特徴ではなく，構文木に重きを置き統語的特徴を抽出する方式で作者を識別する際の有効性を示すことを目的としている．本稿では文章を構成する各要素の出現率を測ることで文脈に依らずテキストデータの特徴を抽出する方式を提案している．Kavuri[2]らは文章ごとの文体の違いを測り機械学習における特徴の，評価，作者予測に焦点をあて作者を予測しそれに基づいた作者プロファイリングを行なっている．本方式によりテキストデータごとの文体の違いを測ることで特徴量を抽出しその中でも重要度の高いものをクラスタリングすることで既存の執筆スタイルではない新たなタイプを見出すことが可能になる．Ding[4]らは文体に基づき作者のアイデンティティ及び社会的特徴を抽出する方式である．提案方式では作者の特徴を抽出し識別する上で複数のジャンルにわたるテキストデータ及びニューラルネットワークが使用されている．本稿では文章の特徴のうち作者ごとにより重要度が高いものを導出することによって各人物に対し特徴づけを行い文章のタイプを可視化する方式を提案している．Graham[5]らはニューラルネットワークを用いて文章の変化する箇所を見つけることで作者が好みとする構文を執筆する支援ツールを提案している．本稿では文章間の機微ではなく一人の作者の性向が異なるテキストデータでどのように変化するのかに着目し方式の提案を進めている．これにより作者個人の潜在的な文体特性の導出を実現している．Argamon[6]らは単語の意味機能の分類に基づき語彙機能を開発することを目的とした方式を提案している．この方式はテキスト分類及び洞察に役立つとされている．本稿では単語をその意味に依らず一つの特徴量として扱うことで文体と同等の作者の性向として分析に用いることを実現している．Pimonova[7]らはロシア語の作者の文体をモデル化するため語彙，形態論，構文の三つの言語レベルで 7 つのモデルを使用することで形態統語的特徴セットを提案している．本稿では作者の文体及び単語特徴を文脈に依らずに且つより直感的にわかりやすく特徴を抽出することで作者ごとの特徴を測っている．Stamatatoes[8]らはギリシャ語のジャンルと作者の観点より文章を分類することを目的とし入力テキストの分析方法を表した上で文体情報を取得するスタイルマーカーを提案している．本稿では作者の観点に依らないあくまで単語及び文体にフォーカスし既存の処理ツールと特徴の計測方式を組み合わせることで新たな特

†武蔵野大学データサイエンス学部　Department of Data Science, Musashino University

徴量を持って文体情報を取得する方式を提案している．伊藤ら[9]は文章の表現体系について文章における品詞の構成比率において名詞が多く動詞，形容詞類は少なく感動詞類はほとんどない文章を要約的文章としている．一方で名詞が少なく，動詞，形容詞類が多く，感動詞類は少ない文章を描写的文章としている．また，MVR については名詞の比率と比較することで先述した表現体系のうちどれに当てはまるのかを判別することが出来る文章指標だとしている．MVR の計算方法については形容詞，形容動詞，副詞，連体詞からなる M のグループの比率を動詞を意味する V のグループで割り 100 をかけることで求められると説明している．名詞比率と MVR のうち名詞比率の方が大きければ要約的文章，反対に MVR の方が大きければ描写的文章であるとしている．また描写的文章であると判定された場合に M のグループと V のグループを比較することで，文章がものごとを形容する描写の多いありさま描写的文章か，物事の動作に関する描写の多い動き描写的かを判別している．金ら[10]は語彙の豊富さを測るモデルとして Guiraud Index を紹介している．Guiraud Index は文章における異語数を総語数の平方根で割ったものであり，値が大きいほど語彙は多様であると言える．

これらの既存研究に対して，本稿では，伝え方タイプを表現するような文章特徴として，文体と使用言語のユーザである作者に特化した特徴をこれまで我々が提案してきた AIME[3]を用いて抽出する手法を提案する．これにより，ユーザである作者に依存した文章特徴を捉えることが可能となり，伝え方タイプ，および，その作者の過去の経験や考えを推定する一歩となりうる．

3. AIME の概要

本節では図 2 に本方式における AIME[3]の概要を示す．

AIME は，AI，機械学習などのブラックボックスモデルの大局的，局所的説明を導出する説明可能な AI(Explainable AI; XAI)手法での一つである．図 2 において，上段は AI，機械学習などのブラックボックスモデルを指す．このブラックボックスモデルは訓練データX, Yを用いることで，特徴$\boldsymbol{x}$を入力すると$\hat{\boldsymbol{y}}$が出力されるモデルを構築する．この AI，機械学習などのブラックボックスモデルの大局的，局所的説明を導出するのに文献[3]の AIME を用いる際には，訓練データのX，訓練データXに対するこの AI，機械学習などのブラックボックスモデルによって導出された出力$\hat{Y}$を用いてブラックボックスモデルの振る舞いを近似する近似逆作要素を導出していた．しかしながら，本稿の目的は，文章の特徴から作者推定をすることではなく，作者に固有な文章中の特徴を導出することである．その際，仮想的に正しい結果のみを導出する AI，機械学習などのブラックボックスモデルがあると想定するならば，そのブラックボックスは訓練データXに対して訓練データの作者ラベルYを導出するはずである．そのため，X, Yを用いて AIME の近似逆作用素$A^{\dagger}$を導出することにより直接的に作者ごとの特徴を導出することが可能となる．

近似逆作要素$A^{\dagger}$自体は大局的な特徴量の重要度そのものを表し，下記の式で導出される．

$$X = A^{\dagger}Y,$$
$$XY^T = A^{\dagger}YY^T,$$
$$XY^T\,(YY^T)^{-1} = A^{\dagger}(YY^T)(YY^T)^{-1},$$
$$A^{\dagger} = XY^{\dagger}\,(YY^T)^{-1} = XY^T$$

ここでY^Tは行列 Y の転置行列を示し，$Y^{\dagger}$は Y のムーアペンローズの一般化逆行列を示す．

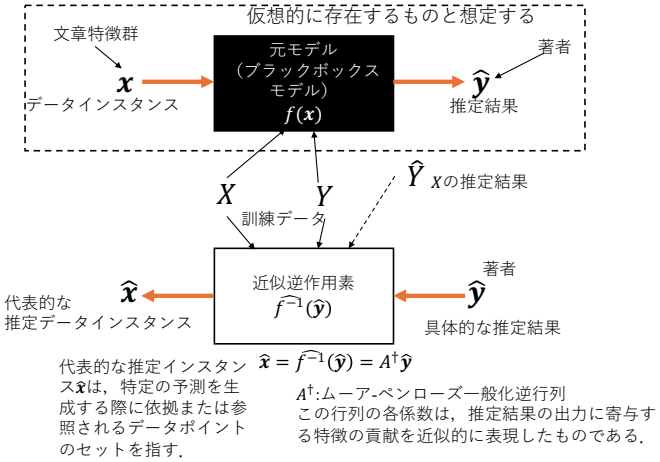

図 2　本方式の目的に適用した AIME の概要

4. 作者の性向を表す語彙・文体の特徴抽出方式

4.1 方式の全体像

図 3 に提案方式の全体像を示す．本方式は形態素解析機能，文体特徴抽出機能，単語特徴抽出機能，作者名正解データ抽出機能，ユーザ別特徴重要度抽出機能，ユーザ別クラスタリング機能からなる．

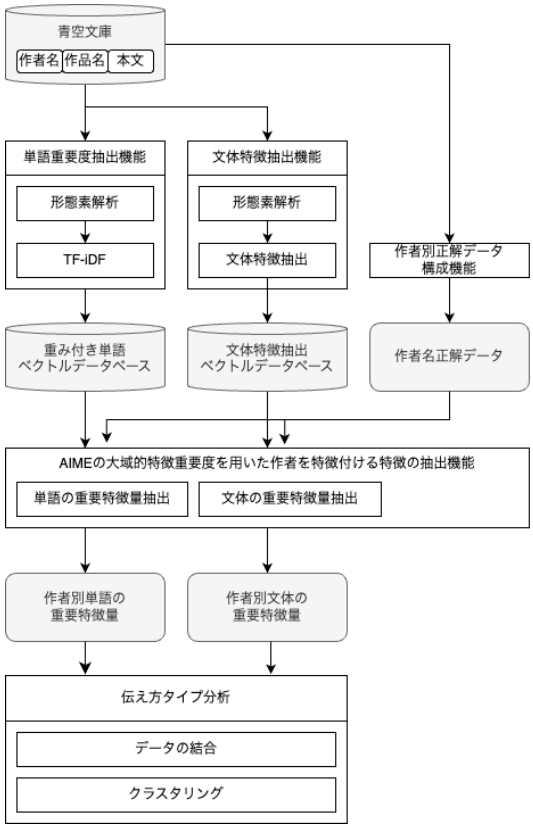

図 3　提案方式の全体像

本方式では各作者が執筆したテキストデータを入力する．入力されたテキストを形態素解析したのち文体特徴及び単語特徴をそれぞれ算出，ユーザ別特徴重要度にてテキストデータの作者を決定付ける要素として貢献度が高いものを導出する．さらにその結果をクラスタリングすることで作者の文章における特徴と伝え方タイプを可視化することを目的とする．

4.2 文体特徴抽出機能

本節では入力したテキストデータを形態素解析した上で文体の特徴を抽出する機能について述べる．

尚，本方式では文体を表 1 に示す 7 項目より計測できるものとする．

表 1 文体特徴抽出に用いる項目

項目	計測手法
品詞出現比率	品詞ごとの出現割合の計測
表現方法の差異	MVR と名詞出現率の比較
使用語彙の豊富さ	Guiraud Index
漢字の数	出現回数の計測
句読点の数	出現回数の計測
文長の平均値	テキストデータごとに平均値をとり正規化
文長の中央値	テキストデータごとに中央値をとり正規化

4.2.1 形態素解析機能

本節では入力されたテキストデータに対して形態素解析を行う機能について述べる．本方式では文体特徴を抽出する上で入力されたテキストデータを単語単位に細分化する必要がある．そこで MeCab と NEologd を用いテキストデータを単語まで分割した上で名詞，動詞，形容詞，数，非自立，接尾，記号，連体詞，助詞，助動詞の 9 つの品詞に割り当てる．

4.2.2 文体特徴抽出

品詞の出現比率は 4.3.1 節にて形態素解析されたテキストデータの単語を名詞，動詞，形容詞，数，非自立，接尾，記号，連体詞，助詞，助動詞の 9 つの品詞のうち当てはまるものを該当品詞の使用量としてカウントし，各品詞の使用量を全体の単語量で割ることで割合を算出している．

文章における表現方法の差異は MVR と名詞の出現比率を比較することで判別する．ここで，本稿での文章における表現方法を要約体，動き描写的文章，ありさま描写的文章の 3 種類に分けられると定義する．

本方式では形容詞と連体詞の出現率を動詞の出現率で割ったものに 100 をかけることで MVR を算出している．名詞の出現率に 100 をかけたものと MVR を比較することでテキストデータが要約的文章か描写的文章かを判別することが出来る．また MVR においても V グループと M グループの出現率を比較することでテキストデータが動作的描写かありさま描写かを判別することが可能になる．

使用語彙の豊富さは Guiraud Index を用いて測る．Guiraud Index は文章における異語数を総語数の平方根で割ったものであり，値が大きいほど語彙は多様であると言える．使用語彙の豊富さを測るには異語数を総語数で割ることで算出する TTR という方法もあるが，計算結果がテキストの長さにより影響されてしまうため，今回は Guiraud Index を用いた．Guiraud Index により語彙に着目したテキストデータにおける作者の表現の多様さを図ることが出来る．

文章中で使用されている漢字の数は正規表現を用いてカウントする．一般に日本語で書かれた文章では仮名と漢字が用いられる．この指標により作者の使用語彙の傾向及び難易度を測ることが可能になる．

句点の数は文章中で使用されている句点をカウントすることで算出した．長く複雑な文章であるほど句点を活用することが必要となる．この指標により作者の文章の複雑さを測ることが可能になる．

本稿では一つのテキストデータあたり構成する文章について句点間の文字数を文長と定義する．文長は作者によって異なり．文長が長いほど文章の構造が複雑になり内容理解に負荷がかかるとされている．本方式では作品ごとに文長の平均値及び中央値を算出，比較することで一文単位での文章理解の難易度を測る指標として用いる．この際平均値と中央値の二つの値をそれぞれ別々の指標としているのは，文章の難易度推定に用いられることの多い平均値のみを使用した場合テキストデータにおける文長のばらつきが考慮できないためである．文長の平均値と文長の中央値は文章中の句点間の文字数を取得することで計算し，その結果は正規化している．

4.3 単語特徴抽出機能

4.3.1 形態素解析機能

本機能では入力されたテキストデータに対して 4.2.1 節で述べたものと同一の手順で形態素解析を行う．

4.3.2 TF-IDF

単語特徴は TF-IDF を用いて使用語彙の出現頻度を算出することで抽出する．4.4.1 節にて形態素解析を行ったテキストデータそれぞれに対し TF-IDF を用いてテキストデータ全体の語彙の出現頻度を測る．TF-IDF の評価によってそれぞれの語彙が各テキストデータにおいてどのように扱われているのかを識別することが可能になる．識別された語彙の扱われ方は各テキストデータの特徴であると考える．

4.4 作者名正解データ抽出機能

入力したテキストデータごとに作者データをつけることで AIME を用いて計算する際の行列 Y のデータを作成する．

4.5 ユーザ別特徴重要度抽出機能

本節では AIME を用いたユーザ別特徴重要度抽出機能について述べる．本方式では文体特徴と単語特徴をテキストデータに現れる作者の特徴として扱っている．作者を決定付ける特徴として貢献度の高いものを導出するにあたって，4.3 節及び 4.4 節の機能で抽出した特徴それぞれに対し AIME を用いて導出をする必要がある．

4.5.1 文体特徴重要度量抽出

AIME による文体特徴重要度の概要を図 4 に示す．文体特徴重要度を抽出する上で 4.2 節にて抽出した文体特徴と 4.5 節にて抽出した作者名正解データを用いる．これは図 5 に示す通り 57(作品)×9(特徴)からなる行列 X_1 と図 6 に示す通り 57(作品)×19(人)からなる行列 Y である．文体特徴を行列 X_1，作者名正解データを行列 Y とした上で 3 節に示した AIME を用いて文体特徴重要度を導出する．

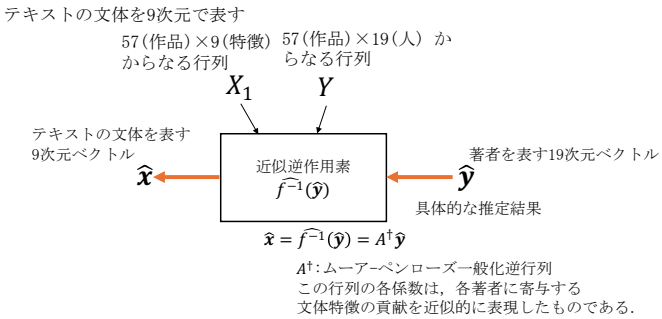

図 4　AIME による文体特徴抽出機能

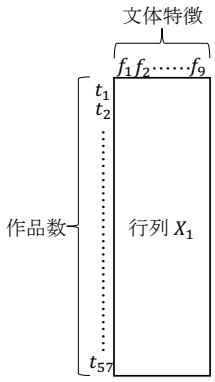

図 5　文体特徴を表す行列 X_1

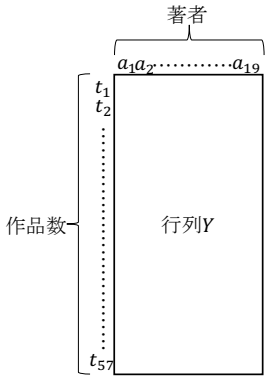

図 6　作者を表す行列 Y

4.5.2 単語特徴重要度抽出機能

AIME による単語特徴重要度の概要を図 7 に示す．単語特徴重要度を抽出する上で 4.3 節にて抽出した単語特徴と 4.5 節にて抽出した作者名正解データを用いる．入力するテキストの特徴$\boldsymbol{x}$は単語特徴を抽出した 16557 次元のベクトルである．これは図 8 に示す通り 57(作品)×16557(特徴)からなる行列 X_2 となる．出力$\hat{\boldsymbol{y}}$は図 5 に示す通り 57(作品)×19(人)からなる行列 Y である．文体特徴を行列 X_2，作者名正解データを行列 Y とした上で 3 節に示した AIME を用いて単語特徴重要度を導出する．

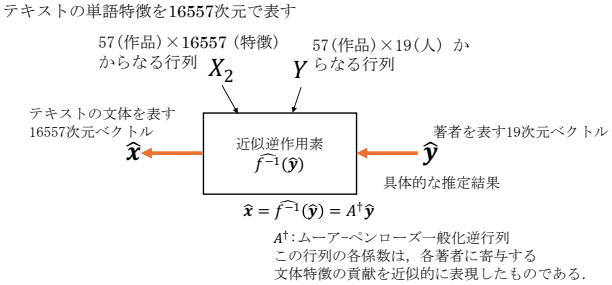

図 7　AIME による文体特徴抽出機能

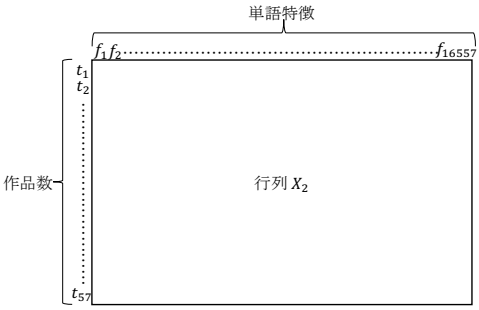

図 8　単語特徴を表す行列 X_2

4.6 ユーザ別クラスタリング機能

本節ではユーザ別クラスタリング機能について述べる．

4.6.1 データ結合

AIME によって抽出した文体特徴及び単語特徴において作者における重要度の高いもののうちそれぞれ上位 9 項目を選出しデータを結合させた．これによりテキストデータあたりの文体特徴重要度及び単語特徴重要度についてクラスタリングすることが可能になる．

4.6.2 クラスタリング機能

本方式の目的は文章における特徴と伝え方タイプを可視化することである．4.6 節より導出された文体及び単語特徴重要度をクラスタリングすることより，作者と他の作者のテキストデータに現れる特徴から文章がどのようなタイプであるかを測ることが可能となる．この際クラスタリングするデータは 4.6.2 節および 4.6.3 節にてより貢献度が高いと計算された上位 9 つのデータをそれぞれ用いる．

5. 実験

本節では 4 節にて述べた提案方式に基づき行った実験について述べる．本実験ではテキストデータに対し特徴抽出，特徴重要度抽出，クラスタリングを順に行った上で，各作者の文章における伝え方の癖を可視化及び考察する．

5.1 実験環境

本実験は Google Colabratory にて行なった．テキストデータは青空文庫より取得した小説の zip ファイルを使用した．小説および作家は表 2 の通りである．青空文庫にて作品が公開されている作家のうち名前がイ音で始まり公開作品が

表 2　実験に使用した作家と作品

作家	作品
生田春月	『幸福が遅く来たなら』『誤植』『象徴の烏賊』
石井研堂	『大利根の大物釣』『元日の釣』『研堂釣規』
石川欣一	『飢えは最善のソースか』『可愛い山』『針の木のいけにえ』
石川三四郎	『蒼馬を見たり』『社会的分業論』『土民生活』
石川啄木	『足跡』『新しい歌の味ひ』『A LETTER FROM PRISON』
石原純	『アインシュタイン教授をわが国に迎えて』『雨粒』『左千夫先生への追憶』
泉鏡花	『愛と婚姻』『悪獣篇』『芥川竜之介氏を弔ふ』
伊丹万作	『映画界手近の問題』『映画の普及力とは』『映画と民族性』
伊藤左千夫	『浅草詣』『市川の桃花』『井戸』
伊藤野枝	『青山菊栄様へ』『新らしき女の道』『新らしき婦人の男性観』
犬養健	『亜剌比亜人エルアフイ』『朧夜』『愚かな父』
犬田卯	『一老人』『おびとき』『錦紗』
井上円了	『欧米各国　政教日記』『甲州郡内妖怪事件取り調べ報告』『西航日録』
伊波普猷	『浦添考』『沖縄人の最大欠点』『「古琉球」改版に際して』
伊庭心猿	『荷風翁の発句』『九月朔日』『桜もち』
今井邦子	『伊那紀行』『誠心院の一夜』『滝』
今村恒夫	『アンチの闘士』『鋼鉄』『山上の歌』
岩野泡鳴	『黒き素船』『札幌の印象』『塩原日記』
岩本素白	『雨の宿』『鰯』『菓子の譜』

5 作以上ある日本人作家 19 人の作品をテキストデータとして選出した．青空文庫における掲載作家別作品リスタのうち zip ファイルからテキストを読み込める 50 音順での上位 3 作品をテキストデータとして使用した．

5.2 実験 1：文体特徴抽出の検証

文体特徴抽出の結果は表 3 に示した通りである．品詞の出現比率について，同一の作家が執筆した作品間では各品詞の出現比率の差は 0.3 以内であることが多かった．文長の平均値と文長の中央値についても二者間の差は 0.001 から 0.004 間にとどまることが多かった．文長の平均値と文長の中央値に 0.004 以上の差がついた場合，句点の数が 0.04 以上となることが多くなっている．表現方法についてはほとんどが簡約体であると判別され，ごく稀に動き描写的文章であると判別されていた．

5.3 実験 2：単語特徴抽出の検証

単語特徴抽出機能の結果は表 4 に示した通りである．ほとんどの作品に対し 0 が振られており，出現している単語に対しては頻度が振られている．

5.4 実験 3：ユーザ別特徴重要度及びそのクラスタリング結果の検証

5.4.1 文体特徴重要度抽出の検証

5.2 節にて抽出した特徴について AIME を用いて各作家を特徴づける上で重要度の高い項目を導出した結果を図 9 に示す．作家によって重要度の高い項目は異なるが品詞の出現比率及び使用語彙の豊富さは多くの作家にとって重要な特徴となっていることがわかる．反対に文体や文長の平均値・中央値に関しては重要度の著しく高い作家が 1 人から 3 人存在しており，特定の作家の文章においては大きな特徴となっている．

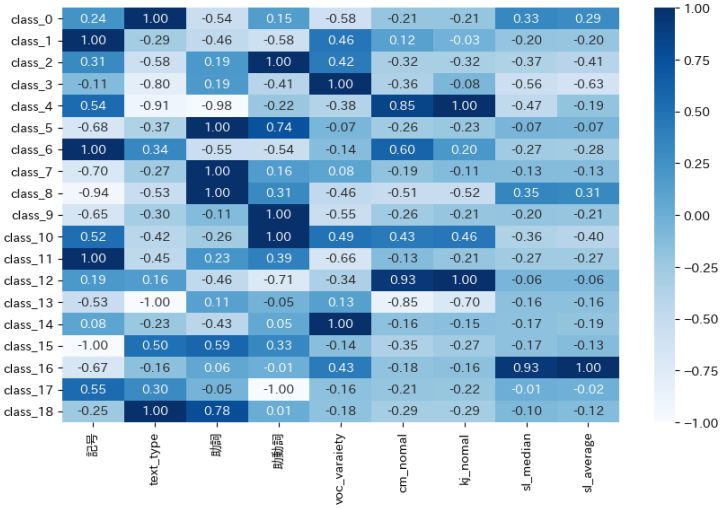

図 9　AIME による文体特徴重要度抽出の結果

5.4.2 単語特徴重要度抽出の検証

5.3 節にて抽出した結果について AIME を用いて各作家を特徴づける上で重要度の高い項目を導出した結果を図 10 に示す．本実験では青空文庫よりテキストデータを取得する上で新字体か旧字体かのフィルターは設けなかった．その結果，同じ意味合いを持つ新字体の単語と旧字体の単語の重要度が高かった．単語特徴はテキストデータ全体で用いられている全単語を対象に抽出作業が行われているため，該当単語が使用されているか否かで重要度の大小が大きく異なり，結果にはばらつきがあった．上位 9 単語であっても大抵の単語の重要度は低くなっていた．

表 3　文体特徴抽出の結果(一部抜粋)

作品	記号	助詞	助動詞	表現方法	使用語彙の豊富さ	句読点	漢字	文長の平均値	文長の中央値
アンチの闘士	0.089631	0.300527	0.084359	0.0	7.958399	0.001623	0.011336	0.163048	0.109890
鋼鉄	0.027950	0.282609	0.086957	0.0	7.123989	0.000361	0.007413	0.259738	0.205128
山上の歌	0.030675	0.303681	0.079755	0.0	7.799771	0.000000	0.007048	1.000000	1.000000
黒き素船	0.172414	0.293103	0.017241	0.0	5.167538	0.003066	0.000532	0.011547	0.015568
札幌の印象	0.159021	0.262997	0.058104	2.0	7.678450	0.021641	0.016223	0.154426	0.146520
塩原日記	0.112766	0.314096	0.080851	0.0	5.974127	0.063300	0.052525	0.045598	0.045788

表 4　単語特徴抽出の結果(一部抜粋)

作品	10 月 15 日	10 月 30 日	…	鼻腔	齎す
アンチの闘士	0.000000	0.000000	…	0.000000	0.000000
鋼鉄	0.000000	0.000000	…	0.000000	0.000000
山上の歌	0.000000	0.000000	…	0.000000	0.000000
黒き素船	0.000000	0.000000	…	0.000000	0.000000
札幌の印象	0.000000	0.000000	…	0.000000	0.000000
塩原日記	0.000000	0.020016	…	0.000000	0.000000

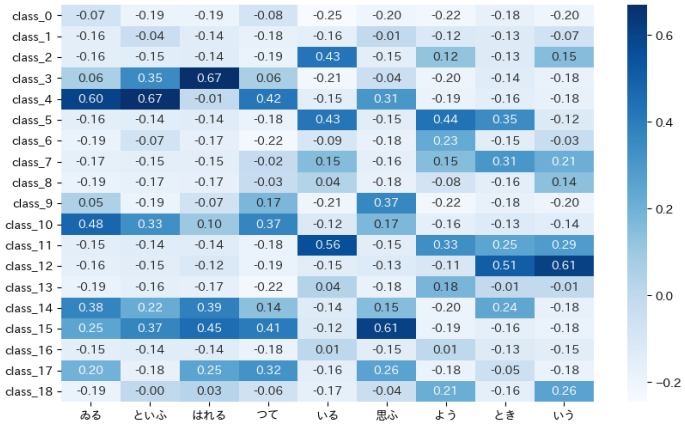

図 10　AIME による単語特徴重要度抽出の結果

5.4.3 クラスタリング結果の検証

5.4.1 節および 5.4.2 節の結果を踏まえ各特徴重要度の上位 9 項目を結合しクラスタリングした．これにより作家を決定づける際より重要度の高い特徴に基づいた伝え方タイプを可視化している．クラスタの数は 3 つ 4 つ 5 つの 3 パターンで行った．以下に示す図 12 はクラスタ数を 5 つに設定し検証をした結果である．凡そ近くにプロットされている作家が同一のクラスタであるとみなされていたが，犬養健と石川啄木，石川欣一と犬田卯は入り組んでクラスタされていた．

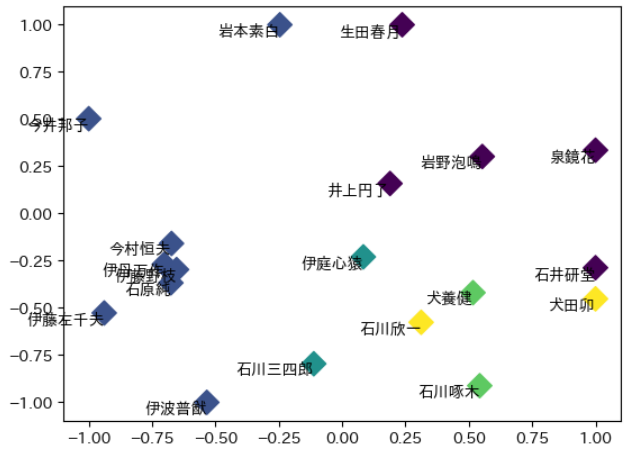

図 12　特徴重要度に基づいたクラスタリング結果

5.5 考察

本実験によりテキストデータの作者ごとに文体および単語の特徴を抽出することが可能であることがわかった．AIME を用いることで作者に対する特徴ごとの重要度も可視化することができた．同一のクラスタとされている作者について AIME によって導出された重要度を合わせて検証する．5.4.2 節の結果を踏まえると同一クラスタの作家は作家を決定づける上で重要度が高いと示された単語が類似していることがわかる．特に 5.4.3 節にて入り組んでクラスタされたと述べた作家について，犬養健と石川啄木は重要度の高い単語がある，といふ，つて，思ふと共通しており，石川欣一と犬田卯においても同様にいる，よう，ときと重要度が高いとされる単語が共通していた．これにより同一の単語に対する重要度がクラスタのされ方に影響すると考えられる．一方で，文体特徴については重要度が高いとされた項目が一致するクラスタも存在したが，多くの場合作家ごとに重要度の示し方は様々で特定の項目とクラスタリングのされ方に関係を見出すことはできなかった．本方式では抽出された特徴から判別した伝え方タイプが類似する作家をクラスタリングにより可視化することを目的として

いたが，結果より文体特徴は伝え方タイプの判別に影響を与えているとは考えにくい．

以上より本方式を用いて文体及び単語特徴を客観的に把握することは可能だが，特に文体特徴に基づいた伝え方タイプを判別し可視化することは困難であると言える．クラスタごとに重要度の一致している項目もあるため文体特徴の項目ごとに特徴抽出方法や求めている特徴が抽出できているのかを見直すことでこの問題は解決できるのではないかと考える．

6. おわりに

本稿では多様な人物が執筆したテキストデータを対象としたその人物の性向を表す語彙・文体特徴の抽出方式を示した．作者ごとに文体，使用語彙それぞれの特徴があり文体特徴抽出機能，単語特徴抽出機能，ユーザ別特徴重要度抽出機能を用いることで客観的に文章の傾向を把握することが実現した．一方でクラスタリングによる伝え方タイプの判別では文体重要度を反映することができなかったためこの問題の解決を今後の課題とする．

今後の展望としては文体特徴重要度の結果を伝え方タイプの判別に反映させた上で文章構成方策の導出方法について検討を進めることが挙げられる．

参考文献

[1] Zhao Chen, Song Wei, Liu Lizhen, Du Chao, Zhao Xinlei, ″Research on Author Identification Based on Deep Syntactic Features,″ in 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Vol. 1, pp. 276-279, (2017).

[2] Kavuri Karunakar, Kavitha M., ″A Stylistic Features Based Approach for Author Profiling,″ in Recent Trends in Communication and Intelligent Systems, Springer Singapore, Singapore, pp. 185-193, (2020).

[3] Nakanishi Takafumi, ″Approximate Inverse Model Explanations (AIME): Unveiling Local and Global Insights in Machine Learning Models,″ in *IEEE Access*, Vol. 11, pp. 101020-101044, (2023).

[4] Ding Steven H. H., Fung Benjamin C. M., Iqbal Farkhund, Cheung William K., ″Learning Stylometric Representations for Authorship Analysis,″ in IEEE Transactions on Cybernetics, Vol. 49, No. 1, pp. 107-121, (2019).

[5] Graham Neil, Hirst Graeme, Marthi Bhaskara, ″Segmenting documents by stylistic character,″ in Natural Language Engineering, Vol. 11, No. 4, pp. 397-415, (2005)

[6] Argamon Shlomo, Whitelaw Casey, Chase Paul, Hota Sobhan Raj, Garg Navendu, Levitan Shlomo, ″Stylistic text classification using functional lexical features,″ in Journal of the American Society for Information Science and Technology, Vol. 58, No. 6, pp. 802-822, (2007).

[7] Pimonova Elena, Durandin Oleg, Malafeev Alexey, ″Authorship Attribution in Russian with New High-Performing and Fully Interpretable Morpho-Syntactic Features,″ in Analysis of Images, Social Networks and Texts, Springer International Publishing, Cham, pp. 193-204, (2019).

[8] Stamatatos Efstathios, Fakotakis Nikos, Kokkinakis George, ″Automatic Text Categorization in Terms of Genre and Author,″ in Computational Linguistics, Vol. 26, No. 4, pp. 471-495, (2000).

[9] 計量国語学会(編集)，荻野綱男，伊藤雅光，丸山直子，長谷川守寿，荻野紫穂(編)，″データで学ぶ日本語学入門,″ 朝倉書店，(2017).

[10] 計量国語学会(編集)，伊藤雅光，土屋信一，大島資生，長谷川守寿，荻野紫穂，山崎誠，荻野綱男，横山詔一，田中ゆかり(編)，″計量国語学事典,″ 朝倉書店，(2010).