テキストマイニングによる筆者識別の 正確性ならびに判定手続きの標準化

財 津 亘*, 金 明 哲**

Accuracy and Standardized Judgment Procedures for Author Identification by Text Mining

Wataru ZAITSU* and Mingzhe JIN**

This study examined the accuracy for author identification by text mining. We conducted 16 analyses (four writing styles $\times$ four multivariate analyses) across texts of 100 Bloggers, written by approximately 1,000 characters. Specifically, we conducted (1) principal components analysis, (2) correspondence analysis, (3) multi-dimensional scaling, and (4) hierarchical cluster analysis on each writing style: (1) rate of usage of non-independent words, (2) bigram of parts-of-speech, (3) bigram of postpositional particles, and (4) positioning of commas. We obtained high accuracy: 100% on sensitivity and 95.1% on specificity. Furthermore, the results showed no effects of age and gender against accuracy for author identification.

Key words: author identification, text mining, multivariate analysis

キーワード: 筆者識別, テキストマイニング, 多変量データ解析

1. は じ め に

犯罪捜査の現場では、脅迫や名誉毀損事件で用いられた印字文書や電子メールなどの文章が、誰によって記載されたのかが重要となる。しかし、それらの文章は手書き文字ではないために、従来の筆跡鑑定による筆者の識別は不可能であった。筆者を識別するもう一つの有効な手法として、文章上の文体的特徴を計量分析する手法が挙げられる。文体的特徴とは、読点の打ち方や文の長さといった文章上の特徴のことである。計量的文体分析による筆者識別法は、諸外国では 19 世紀半ばに初期の研究がみられ (金, 2009; 村上, 2016), 英

国では犯罪捜査における活用も見受けられる (Grant, 2013)。一方、我が国では、計量文献学や計量文体学分野において、日記や文学作品を対象に研究が行われてきた経緯はあるものの、犯罪捜査への活用例は少ない。また、近年コンピュータの発達・普及にともない、計量文体学における手法と知見を援用・拡張する形で、テキストマイニングと呼ばれる手法が出現し、注目されている (金, 2016)。金 (2009) によると、テキストとは、文章のように記号列が何らかの規則に従い並べられた集合を意味し、テキストマイニングとは、それらの記号列をある単位 (単語など) に分割し、それらの出現頻度を定量的に統計解析する手法の総称である。

テキストマイニングによる筆者識別が可能となる根拠は、筆跡鑑定 (高澤, 1998) と同様で、文章を書くという行動における個性が個人識別の手がかりとなっている。つまりは、各個人によって表出される文体的特徴に「個人差」と「個人内恒常性」が存在することから筆者の識別が可能となる (Grant, 2013)。

具体的な例として、名誉毀損事件などの犯罪に使われ、かつ筆者が不明である文章 (テキスト形式の文章データを扱った電子ファイルの意味を含めて、以下、疑

*富山県警察本部刑事部科学捜査研究所
(Forensic Science Laboratory, Toyama Prefectural Police Headquarters)

連絡先: 〒930-8570 富山県富山市新総曲輪 1 番 7 号
E-mail: wataru0112csi@yahoo.co.jp

**同志社大学文化情報学部

(Faculty of Culture and Information Science, Doshisha University)

連絡先: 〒610-0394 京都府京田辺市多々羅都谷 1-3
E-mail: mjin@mail.doshisha.ac.jp

問テキストとする)が、犯罪捜査上に浮上した被疑者によって記載されたものか否かを判定するとしよう。その場合、被疑者が記載したことが明確である文章(疑問テキストと照らし合わせる意味で、以下、対照テキストとする)を収集する。また着目する文体的特徴において個人差を検討するために、被疑者とは別人の文章を複数人分用意する(以下、無関係テキストとする)。財津・金(2017a)ならびに財津・金(2017b)を例に挙げると、上記の疑問・対照・無関係テキストにおける非自立語の使用率、品詞の bigram、助詞の bigram、読点の打ち方、漢字や仮名などの使用率、文の長さといった文体的特徴に着目し、それぞれの文体的特徴について、主成分分析(PCA, principal components analysis)や対応分析(CA, correspondence analysis)、多次元尺度法(MDS, multi-dimensional scaling)、階層的クラスター分析(HCA, hierarchical cluster analysis)といった多変量データ解析を実施している。これら多変量データ解析では、2次元上にそれぞれのテキストを布置するもしくは樹状図(デンドログラム)としてテキスト間の類似度を表現する形で分析結果が算出される。しかし、従来は、これらの分析結果を、視覚的にみて総合的に筆者を識別していたため、分析者によって判定結果が異なる可能性があった。そのために、財津・金(2017a)ならびに財津・金(2017b)は、これらの分析結果(疑問・対照・無関係テキストの布置関係やデンドログラムのクラスター関係)に対して、後述するスコアリングルールに沿って得点を付与し、その得点の合算値を基に疑問・対照テキストが同一人によるものか否かを判定するシステムを採用し、その妥当性を示した。

なお、テキストマイニングによる筆者識別研究では、ランダム・フォレスト(Breiman, 2001)やサポートベクターマシン(Vapnik, 1995)といった分類器を用いて筆者を判別するものが数多く存在する。しかし、法廷では分析結果に関する説明責任がともなうことから、テキスト間の類似度を基に、次元上にテキストを布置するまたはクラスターとしてテキストをまとめるといった視覚的に理解しやすい解析手法を採用する方が望ましいといえる。

また、米国の法廷においては、分析手法が証拠能力を有するか否かを判断するための評価基準にドーバート基準といったものがある(勝又, 2008)。ドーバート基準とは、①その手法が、実証的に検証可能であること、②理論や技術が、ピア・レビューや出版されているこ

と、③結果を評価するための誤差率や標準的な手続きが明確であること、④適切な専門的分野で一般的に承認されていることの4つである。したがって、テキストマイニングを用いた筆者識別法が、法廷において証拠能力を認められるためにも、③の誤差率にかかる筆者識別の精度ないし正確性の検証が求められる。例えば、警察で行われる鑑定の一つに、ポリグラフ検査があるが、小川・松田・常岡(2013)は、このポリグラフ検査の正確性に関する実証研究を行い、判定不能率が20%で、判定不能を除外した場合の感度が86%、特異度が95%といった精度を得ている(感度・特異度については後述)。以上から、テキストマイニングによる筆者識別についても同様に、正確性に関する検証が必要といえるが、現在までの同種研究(金, 2014; 三品・松田, 2013)の多くが、分類器による識別精度の検証であり、本手法の正確性を検証した研究は少ない。財津・金(2017a)では、PCAやCA、MDSを扱い、財津・金(2017b)では、HCAを用いて正確性を検証しているが、両研究ともに筆者10名とサンプル数が少なかった。このことから、本研究ではサンプル数を増やし、100名が記載したブログの文章を対象に、テキストマイニングによる筆者識別の正確性を検証することを目的とした。加えて、財津・金(2017a)や財津・金(2017b)では、30代男性のテキストに限定していることから、本研究では、5つの年齢層(20代から60代)の男女それぞれ10名の全100名の筆者を対象に、年齢や性別の正確性への影響も合わせて検討した。

科学的鑑定手法は、高い正確性を有するとともに、再現性といった科学原則に則っている必要がある。着目する文体的特徴や用いる多変量データ解析を定め、かつ財津・金(2017a)や財津・金(2017b)のように、複数の多変量データ解析の分析結果に対して得点を付与し、その合算値を基に筆者を判定することが可能となれば、分析者が異なっても、同一手順を踏むことで同一の判定結果が得られる。これは、科学的鑑定の要諦である再現性を有し、かつドーバート基準の要件の一つである判定手続きの標準化も可能になることを意味する。このことから、本研究では、この点についても、正確性の検証を踏まえた上で考察したい。

2. 方 法

2.1. サンプル

本研究では、インターネットサイト「にはんブログ村

(<http://diary.blogmura.com/>)」内の、自称 20 代から 60 代の男女 50 名ずつ計 100 名が記載した 100 のブログから、文章を抽出した (5 つの年齢層 × 2 つの性別の 10 グループそれぞれで 10 名ずつを抽出)。ブログの選定は、内容に依存せず、無作為に行い、20 代男性であれば、「その他 20 代男性日記」を対象に、文章を抽出した時点 (平成 28 年 3 月 17 日) のランキング 1 位から順に選定した。なお、各ブログの話題内容は、個人の趣味や日記などであったことから、同一ブログの文章は、同一筆者によって記載されたものとみなした。

なお、本研究は、インターネットサイト上のブログから文章を収集し、統計解析したものであるが、著作権については、平成 21 年著作権法改正にかかる第 47 条の 7 (情報解析のための複製等) に従っており、何ら問題がない。また、テキストデータについては、日本心理学会の倫理規定に則り、適切に管理している。

2.2. テキストの作成

上記 100 名の文章 (各筆者およそ文字数 4,000 文字程度) を基に、1 名につき 1 テキストファイル (以下、テキストとする) を作成した。次に、それら 100 名の 100 テキストの冒頭 1,000 文字程度を疑問テキストとして分割した。また、残り 3,000 文字程度を対照・無関係テキストとし、筆者ごとに文単位で繰り返しのないようランダムサンプリングを実施し、対照・無関係テキストを各筆者 3 テキストファイルに分割した。したがって、作成されたテキスト数は、疑問テキストで 100、対照・無関係テキストで 300 となった。文の区切りは、1,000 文字であれば、1,000 文字を超えた最初の句点までとした。なお、それぞれのテキスト種別 (疑問・対照・無関係) は機械的かつ無作為に分割したため、テキスト種別 (疑問・対照・無関係) 間における内容の類似性については考慮していない。

2.3. 分析デザインとテキストの組合せ

本研究では、財津・金 (2017a) や財津・金 (2017b) と同様に、筆者組合せ条件「同一人」と「別人」を設定した。

筆者組合せ条件「同一人」であれば、疑問テキストと対照テキストに同一人のテキストを組合せた (例、筆者 A の疑問テキスト 1 つ、筆者 A の対照テキスト 3 つ)。無関係テキストについては、疑問・対照テキストと異なる筆者のテキスト (4 名分) を、疑問・対照テキストの筆者と同じ年齢層・性別グループからランダムに抽出し、

割り当てた (各筆者 3 テキストの計 12 テキスト)。これら 16 テキストを 1 分析ユニット (小) と定義した。

筆者組合せ条件「別人」の、対照・無関係テキストの組合せは、筆者組合せ条件「同一人」と同じ組合せとした。疑問テキストについては、対照・無関係テキストと同じ年齢層・性別グループ (疑問テキストの筆者を除く) から、繰り返しのないランダムサンプリングを実施して決定した。

2.4. 分析ユニット (大)

上記 1 分析ユニット (小) に対して、後述する 16 分析 (4 文体的特徴 × 4 多変量データ解析) を実施し、この単位を 1 分析ユニット (大) とした。これは、財津・金 (2017a) や財津・金 (2017b) の「1 分析ユニット」に相当し、実務において筆者識別鑑定を 1 回行うことを想定した単位である。本研究では、筆者組合せ条件「同一人」「別人」それぞれで 100 分析ユニット (大)、計 200 分析ユニット (大) の得点を分析対象とした。

2.5. 文体的特徴

以下の 4 つの文体的特徴に着目し、分析を行った。これらの文体的特徴は、財津・金 (2017a) や財津・金 (2017b) において、筆者の識別力が高かった特徴であったことに加えて、内容に依存しないといった利点が挙げられる。一方、漢字・仮名などの使用率や文の長さ、動詞の長さの識別力は低かったことから、本研究では用いていない。本研究では、各テキスト内の文体的特徴の出現頻度を算出したが、度数は相対度数 (各セルの度数をテキストごとで合算した総度数で割った値) に変換して扱った。

(1) 非自立語の使用率

副詞、助動詞、助詞、(内容を表す自立語を除く) 動詞、(自立語除外、代名詞は含む) 名詞、(自立語除外) の形容詞、連体詞、感動詞、記号、接続詞、接頭詞の単語の出現頻度を分析の対象とした。また、後述のソフトウェアを用いることで、各品詞を第 3 階層まで分類できるが、情報量が最も多くなる利点から、本研究では以降の文体的特徴に関する分析も含めて、品詞の第 3 階層までの情報を用いた。

非自立語の使用率の具体例を挙げると、それぞれのテキスト内の「と(助詞_格助詞_一般)」, 「そこ(名詞_代名詞_一般)」, 「,(記号_読点)」などの変数の出現頻度を集計した。

(2) 品詞の bigram

品詞の bigram とは、文章を「品詞」に分解した後、隣接する 2 つの品詞を変数として、テキスト内のそれらの変数の出現頻度を集計する方法である。

例としては、「助詞(格助詞__連語)+名詞(代名詞__一般)」、「名詞(代名詞__一般)+動詞(非自立)」などを変数とし、テキスト内の各変数の出現頻度を集計した。

(3) 助詞の bigram

上記品詞の bigram とは異なり、「助詞の単語」を対象にした bigram である。

例を挙げると、「に(助詞__格助詞__一般)+の(助詞__連体化)」、「の(助詞__連体化)+か(助詞__終助詞)」を変数に、テキスト内の出現頻度を集計した。なお、データセットにおける各変数の合計度数が 5 以下の場合は、そのような変数をすべて変数 others として 1 つにまとめた上で分析を行った。

(4) 読点の打ち方(単語)

読点の打ち方は、従来から筆者識別に有効であると指摘されている(金, 1994)。ただし、金(1994)では、読点の前の「文字」または「品詞」を対象としているが、本研究では、財津・金(2017a)や財津・金(2017b)と同様の、読点の前の「単語」に着目した。その理由としては、「単語」は、「文字」と「品詞」の両情報を含むために情報量が豊富で、かつ筆者の識別力の高さが示されているためである。

以上、それぞれのテキスト内における、読点の前の「に(助詞__格助詞__一般)」や「し(動詞__自立)」、「が(助詞__接続助詞)」といった単語の度数を算出した。読点の打ち方についても、助詞の bigram と同様の方法で、低頻度の変数を変数 others としてまとめて扱った。

2.6. 多変量データ解析

上記 4 つの文体的特徴それぞれに対して、以下の 4 つの多変量データ解析を実施した。

(1) 主成分分析(PCA)

連続データで構成されたデータセットを基に、情報の損失を抑え、複数の変数をより少ない変数(2 ないし 3 次元)に合成する多変量データ解析の一種である。本研究では、相関係数行列に基づき、2 次元解に沿ってテキストごとに付与される主成分得点を基に、各テキストを布置した。

(2) 対応分析(CA)

分割表など度数データを基に、複数の変数を合成し、低次元に圧縮する多変量データ解析の一種で、2 次元解

に基づいて付与される各テキストの得点によって、それぞれのテキストを布置した。

(3) 多次元尺度法(MDS)

複数のテキストの類似度(距離)を基に、それぞれのテキストを 2 次元上に布置した。本研究では、距離に Kullback-Leibler 距離(Jin & Jiang, 2013, 以下, KL 距離とする)を用いて、古典の多次元尺度法を実施した。

(4) 階層的クラスター分析(HCA)

複数のテキストの類似度(距離)を基に分析を行う点で MDS に類似しているものの、結果を樹形図(デンドログラム)として表現し、類似した個体間が同一クラスター(群)を形成するといった点で異なる。距離は MDS と同様に、KL 距離を用いたが、クラスター形成のアルゴリズムには Ward 法を採用した。

MDS と HCA には、同一の KL 距離を扱ったが、異なるアルゴリズムを基に分析されるため、本研究では異なる解析手法として両分析を実施した。また、2 次元上に個体を布置する点で類似した手法である PCA や CA、MDS を実施することで得られる増分妥当性についてはすでに先行研究で確認されている(財津・金, 2017a)。そこで、本研究では、これに HCA を加えた 4 手法間の相関係数や各手法の AUC (area under the curve) を算出することでその増分妥当性についても確認した。なお、AUC とは、ROC (receiver characteristic curve) 分析によって算出される、2 分類の識別力に関する評価指標である(Macmillan & Creelman, 2004)。AUC は、陽性と陰性を分けるカットオフ値の設定にともなって変化する感度と特異度によって描かれる曲線下面積のことで、陽性群と陰性群を全く識別できない場合の 0.5 から、完全に識別できる場合の 1 までの値を範囲にとる。

2.7. ソフトウェア

本分析には、多言語テキストマイニングツール「MT-MineR (ver.5.3)」を使用した(石田・金, 2012)。文章を分割し、品詞情報を付与するといった形態素解析には、「茶筌(ChaSen)」を使用した(奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座[松本研究室], 2011)。統計解析では、統計解析フリーソフトウェア「R (ver.3.2.3)」を用いた(R Development Core Team, 2015)。AUC の算出には、R のパッケージ「pROC (<https://cran.r-project.org/web/packages/pROC/pROC.pdf>)」を使

図 1. 2次元上のテキスト布置関係を基にしたスコアリングルール

*財津・金 (2017a) を参考に作成，*S：疑問テキスト，C：対照テキスト，I：無関係テキスト

用した。

2.8. スコアリングルール

分析ユニット（小）の16テキストについて，上記4つの文体的特徴に着目し，4つの多変量データ解析を実施したことから，16の分析結果（12の2次元解，4つのデンドログラム）が算出されることとなる。本研究では，それらの分析結果に対して，財津・金 (2017a) や財津・金 (2017b) と同様のスコアリングルールを適用し，それぞれの分析結果に得点を付与した。スコアリングルールを改めて次に記載する。

(1) 2次元上のテキスト布置関係を基にした得点付与の方法

図1に，PCA，CA，MDSで得られた各テキストの布置関係と付与する得点について記載した。このルールは，財津・金 (2017a) を基にしている。まず，筆者ごとに対照・無関係テキストを結び，三角形を作成する。この三角形は，個人内恒常性の範囲を意味する概念といえる。そこで，それら対照テキストの三角形と無関係テキストの三角形が重なった場合は，個人差がないものと判断し，0点とした。また，両種類のテキストに重複がないことを前提に，疑問テキストが，対照テキストの三角形内に布置した場合に2点，対照テキスト外に布置し，かつ無関係テキストに比べて，対照テキストに最も近接した場合は1点とした。逆に，疑問テキストが，対照テキストに比べて，無関係テキストに最も近接した場合は-1点とした。

(2) デンドログラムのテキスト位置関係を基にした得点付与の方法

図2は，HCAで得られるデンドログラムに基づくスコアリングルールを示した。本ルールは，財津・金 (2017b) に則った。第1に，3つの対照テキストと無関係テキストが混在した場合に0点とした。第2に，3つの対照テキストのみでクラスターを形成し，かつ対照テキストのクラスターが，疑問テキストを包含した場合に2点，疑問テキストが対照テキストのクラスターに最も近接した場合に1点，その他の場合に-1点と設定した。

2.9. 分析手続き

以上で説明したとおり，本研究では，1つの疑問テキスト，3つの対照テキスト，12の無関係テキストで構成される1分析ユニット（小）について，4文体的特徴×4多変量データ解析の16分析を実施し，それを1分析ユニット（大）とした。また，筆者組合せ条件「同一人」で100分析ユニット（大）を，筆者組合せ条件「別人」で別の100分析ユニット（大）を扱った。したがって，多変量データ解析の実施回数は，4文体的特徴×4多変量データ解析×100分析ユニット（大）×2筆者組合せ条件の3,200に及んだ。

上記分析後，3,200すべての分析結果に得点を付与し，分析ユニット（小）内の16分析における得点を合算した（分析ユニット（大）を単位とした得点）。続いて，判定にかかる閾値を設定し，全200分析ユニット

図 2. デンドログラムのテキスト位置関係を基にしたスコアリングルール

*財津・金 (2017b) を参考に作成, *S: 疑問テキスト, C: 対照テキスト, I: 無関係テキスト

図 3. 筆者組合せ別 (同一人・別人), 全 200 分析ユニット (大) の得点分布

(大) の得点と閾値を基に, 判定不能率を算出した. また, 「判定不能」を除外した上で, 感度 (筆者組合せ条件「同一人」で正しく「同一人」と判定した割合) と特異度 (筆者組合せ条件「別人」で正しく「別人」と判定した割合), 判定不能率を算出した. 判定基準 (閾値) については, ポリグラフ検査の正確性に関する先行研究 (小川他, 2013) に比べて, より厳格な基準を設定し, 95% 以上の特異度で, かつ感度がより高く, 判定不能率がより少ない判定基準を探索的に設定した. また, ROC 分析による AUC に加えて, 効果量 (Cohen, 1988) を算出し, これらの 95% 信頼区間を求めた.

本研究では, 年齢層や性別の違いによる正確性の違いを検討すべく, 全 200 分析ユニット (大) の得点を基に, 2 (筆者組合せ, between 要因) $\times$ 5 (年齢層, between 要因) $\times$ 2 (性別, between 要因) の 3 要因分散分析を実施した.

続いて, 多変量データ解析間の相関係数と多変量デー

タ解析ごとの AUC を算出した. 加えて, 3 多変量データ解析 (PCA・CA・MDS) の AUC, それに HCA を加えた 4 多変量データ解析の AUC を算出し, 両者の AUC を基に z 検定を実施し, 4 つの多変量データ解析を行う意義 (増分妥当性) を検討した.

3. 結 果

3.1. 筆者識別に関する正確性

図 3 に, 筆者組合せ条件別の得点 (全 200 分析ユニット) 分布を示した. 図 3 によると, 条件別で得点が比較的分離した. 筆者組合せ条件「同一人」の平均得点は 9.49 点 (SD 0.96), 中央値は 9.50 点であった. 一方, 筆者組合せ条件「別人」の平均得点は -7.44 点 (SD 1.58), 中央値は -9.0 点であった.

本研究では, 閾値を絶対値 4 と設定し, 得点の合算値 (分析ユニット (大) の得点) が -4 点から +4 点の範

表 1. 各多変量データ解析間の相関係数

	PCA	MDS	CA	HCA
PCA	—	0.80**	0.84**	0.76**
MDS		—	0.87**	0.84**
CA			—	0.83**
HCA				—

** $p < .01$

囲を「判定不能」とした（図 3 の灰色の範囲）。この場合で、明確に結論を出すことができた事例は、200 分析中 167 分析（83%）、判定不能率は「同一人」群で 14%、「別人」群で 19%であった。また、「判定不能」を除いた上での、感度は 100%（「同一人」群 86 名中 86 名）、特異度 95.1%（「別人」群 81 名中 77 名）であった。

ROC 分析による AUC（95%信頼区間）は、0.98 [0.97, 1.00] となり、swets (1988) の評価基準「高 (0.9～1.0)」に相当する値であった。他方、効果量 d （95%信頼区間）も、3.27 [2.85, 3.70] であり、Cohen (1988) の評価基準「大 (0.8)」を大幅に超える結果が得られた。

3.2. 年齢ならびに性別の正確性への影響

2（筆者組合せ）× 5（年齢層）× 2（性別）の 3 要因（between）分散分析を実施したところ、筆者組合せの主効果のみが有意であった（ $F(1, 180) = 537.4$, $p < .000$ ）。したがって、年齢や性別による筆者識別の正確性への影響はみられなかったといえる。

3.3. 各多変量データ解析間の相関と増分妥当性

分析ユニット（大）を単位に、4 つの文体的特徴の得点を合算して算出した各多変量データ解析（PCA, MDS, CA, HCA）の得点を基に、Pearson の積率相関係数を算出した（表 1）。表 1 によると、相関係数は 0.76 から 0.87 であり、すべての相関係数において有意差がみられた。なお、財津・金 (2017a) の基準文字数 1,000 文字の相関係数 (0.78～0.89) と比べるとほぼ同程度であった。同一距離（KL 距離）を用いた HCA と MDS の相関係数に関しては、他の相関係数に比べて特段高いとはいえなかった。

また、多変量データ解析ごとの AUC（95%信頼区間）を算出したところ、正確性の高い順に、HCA (0.97 [0.94, 0.99])、MDS (0.96 [0.94, 0.98])、CA (0.95 [0.92, 0.98])、PCA (0.94 [0.91, 0.97]) の AUC

が高かった。3 つの多変量データ解析（PCA・MDS・CA）で、0.97 [0.95, 0.99] の AUC が得られ、それに HCA を加えた 4 つの多変量データ解析において、0.98 [0.97, 1.00] の AUC が得られた。両者の AUC に関して、 z 検定を実施した結果、有意差がみられた（ $z = 2.70$, $p < .006$ ）。このことから、4 つの多変量データ解析を実施することの意義（増分妥当性）があることが示された。

4. 考 察

4.1. 筆者識別の正確性と増分妥当性

100 名のブログ筆者を基にした本研究によると、感度で 100%、特異度で 95.1%といった筆者識別の正確性が得られた。加えて、AUC ならびに効果量 d の値も、非常に高い数値を示した。一概に比較はできないが、ポリグラフ検査の正確性研究（小川他, 2013）における 1 質問表の感度（86%）と特異度（95%）に比べても、本手法の正確性の高さがうかがえる。ただし、本研究のテキストはすべて 1,000 文字程度の文字数を扱ったものであるが、財津・金 (2017a) や財津・金 (2017b) の指摘どおり、文字数は筆者の識別力に多大な影響を及ぼす。したがって、1,000 文字未満であれば、感度・特異度はより低下もしくは判定不能率が増加する可能性があり、逆に文字数が多いほど、正確性が増すものといえる。犯罪捜査場面で扱う文書などの文字数は、罪種にもよるが、数千や 1 万文字以上に至る場合もあることから（財津・金, 2015）、1,000 文字程度を対象とした本研究で得られた結果は、ある程度妥当性を有するものと考えられる。その他、筆者の識別力に影響すると思われる要因として、文章作成時の筆者の作為性が挙げられる。犯罪場面では、証拠を残さないように犯人が自己の文章を意図的に偽る可能性がある（Juola, 2006）。したがって、自然な状態で記載された文章と比較すると、意図的に文章上の文体的特徴を偽ることで、正確性が低下する可能性がありえる。このことから、今後この作為性の影響について検討する必要がある。

本研究では、年齢と性別による筆者の識別力への影響はみられなかった。犯罪捜査においては、あらゆる年齢や性別の被疑者を捜査対象とするが、これらの要因に影響されないということは、分析対象者を限定する必要がないことを意味することから、実務上有益といえよう。

財津・金 (2017a) で示された PCA・CA・MDS の増分妥当性に加えて, HCA を加えることでさらに正確性が増すことも合わせて示された. つまりは, 3 つの多変量データ解析を実施する場合に比べて, 4 つの多変量データ解析を実施することで, それぞれの文体的特徴に関する分析結果の信頼性を高めるものと思われる.

4.2. 筆者識別における判定手続きの標準化

続いて, 判定手続きの標準化について考察したい. 判定手続きの標準化とは, 常に同一の文体的特徴に着目し, かつ特定の多変量データ解析を行うことで, たとえ異なる分析者または異なる分析時期であっても, 同じ分析結果 (テキストの次元上の布置関係など) を得ることができ, それらの分析結果を総合することで同じ判定結果 (疑問・対照テキストの筆者は, 「同一人」もしくは「別人」によって記載されたものといった判定) を導き出すことができることを意味する. しかし, 従来は, 同一の分析結果が算出できたとしても, 複数の分析結果を見た目で判断することで筆者を判定していたことから, 分析者もしくは分析時期が異なることで, 筆者に関する判定結果が異なる可能性を否定できなかった. 本研究は, 財津・金 (2017a) や財津・金 (2017b) の筆者 10 名に比べて, 100 名というサンプル数を扱うことで, このような高い正確性を得ることができた. これは, 財津・金 (2017a) や財津・金 (2017b) に比べると, より一段と判定手続きの標準化が可能であることを示したものと見える. 特に, この標準化において重要となるのが得点の閾値である. 本研究では, 絶対値 4 以内の場合は「判定不能」と設定した. 閾値が明確であれば, 新たに筆者識別を実施するに際して, 合算値が 9 点であったことから「同一人」と判定する, もしくは -8 点であったから「別人」と判定するといったことが可能となってくる. ただし, 留意すべき点としては, 本研究で示した閾値は, ブログ上の文章であり, 他の媒体 (例, 電子メールや手書きの文書など) における文章の場合は, 閾値が異なることもありえる. また, 財津・金 (2017a) のとおり, 閾値は扱う文字数が少なくなるほど広がる傾向にあり, 判定不能率が上昇するとされる. したがって, 異なる媒体の文章や異なる文字数の文章を収集し, 本研究と同様の手続きを採用することで, 得点分布を比較し, 媒体の影響を検討する必要がある.

4.3. 今後の課題と展望

今後の課題としては, 前述のとおり, 作為を持って記載された文章の影響や媒体の違いによる影響を検討し, 閾値をより明確にする必要がある他, 財津・金 (2017a) も指摘しているように, 対照・無関係テキストの選定基準の明確化や対照・無関係テキストにおいて実施される文単位のランダムサンプリングの妥当性も検証する必要がある. また, 本研究では, 無関係テキストの筆者は 4 名と設定したが, この無関係テキストの筆者数の影響も検討する必要があるだろう.

指紋や DNA 型, 筆跡などの証拠となりえるものはないが, 印字された文書や電子メールの文章のみが存在するといった事件は世の中に散見される. 従来着手さえできなかった事件が, 本手法による鑑定結果を端緒として, 事件解決の糸口となりえるのである. したがって, 上記のとおり, 課題は山積みであるものの, 本手法は今後の犯罪捜査において有用な分析技術となることは間違いない. ただし, 我が国では, 本手法に対する捜査機関の認知度が非常に低いのも現状である. このことから, 上記課題を克服しつつ, 本手法の周知に努める必要もあろう.

付 記

本研究の一部は, 日本法科学技術学会第 22 回学術集会において発表したものである.

参 考 文 献

- Breiman, L. (2001). Random forest. *Machine Learning*, **45**, 1, 5–32.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Lawrence Erlbaum.
- Grant T. (2013). TXT4N6: Method, consistency, and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, **21**, 2, 467–494.
- 石田基広・金明哲 (2012). コーパスとテキストマイニング. 共立出版.
- Jin M. and Jiang M. (2013). Text clustering on authorship attribution based on features of punctuation usage in Chinese. *Information*, **16**, 7, 4983–4990.
- Juola P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, **1**, 3, 233–334.
- 勝又義直 (2008). 裁判所における科学鑑定の評価について. 日本法科学技術学会誌, **13**, 1, 1–6.
- 金明哲 (1994). 読点の打ち方と文章の分類. 計量国語学, **19**, 7, 317–330.
- 金明哲 (2009). テキストデータの統計科学入門. 岩波書店.
- 金明哲 (2014). 統合的分類アルゴリズムを用いた文章の

- 書き手の識別. 行動計量学, **41**, 1, 35–46.
- 金明哲 (2016). 計量文献学の基礎研究とその応用. (村上征勝・金明哲・土山玄・上阪彩香 計量文献学の射程. pp.57–114, 勉誠出版).
- Macmillan, N.A. and Creelman, C.D. (2004). *Detection theory: A user's guide (2nd ed)*. Lawrence Erlbaum.
- 三品光平・松田眞一 (2013). 文章の書き手の同定における分類法の精度比較. 南山大学紀要 (アカデミア・情報理工学編), **13**, 35–46.
- 村上征勝 (2016). 計量文献学—文献の新たな研究法—. (村上征勝・金明哲・土山玄・上阪彩香 計量文献学の射程. pp.7–55, 勉誠出版).
- 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 (松本研究室). ChaSen—形態素解析器—, 2011, available from [http://chasen-legacy.osdn.jp/], accessed January 15, 2016.
- 小川時洋・松田いづみ・常岡充子 (2013). 隠匿情報検査の妥当性—記憶検出技法としての正確性の実験的検証—. 日本法科学技術学会誌, **18**, 1, 35–44.
- R Development Core Team (2015). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, available from [https://cran.r-project.org/], accessed January 15, 2016.
- Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 4857, 1285–1293.
- 高澤則美 (1998). 筆跡鑑定. 科学警察研究所報告法科学編, **51**, 2, 43–53.
- Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.
- 財津亘・金明哲 (2015). テキストマイニングを用いた犯罪に関わる文書の筆者識別. 日本法科学技術学会誌, **20**, 1, 1–14.
- 財津亘・金明哲 (2017a). テキストマイニングを用いた筆者識別へのスコアリング導入—文字数やテキスト数, 文体的特徴が得点分布に及ぼす影響—. 日本法科学技術学会誌, **22**, 2, 91–108.
- 財津亘・金明哲 (2017b). 階層的クラスター分析結果にスコアリングを導入したテキストマイニングによる筆者識別. 科学警察研究所報告, **66**, 2, 75–81.
- (2017年6月2日受理, 2017年8月2日採択)
(この間審査2回・審査期間合計35日)