Estimating an Author's Age Group by Machine Learning for Offender Profiling

Wataru ZAITSU*, Mingzhe JIN**

(Received March 6, 2018)

It is recommended that criminal investigators use offender-profiling analysis to arrest offenders early and to conduct the criminal investigation efficiently. In offender profiling, analyst estimates an offender's gender, age, or job based on the analysis of the crime scene. In case, there are only printed documents or e-mails available, and therefore it is difficult for analysts to estimate the offender's characteristics. The purpose of this study was to estimate the text authors' age group by using random forests and support vector machines on the basis of stylometric features of texts. The results showed that there were statistical significances among five age groups with next stylometric features of texts on a 100 blog; the frequency of (1) a noun, (2) a binding particle 「は」 just before commas, (3)「ずっと (an adverb)」, and (4) bigram of parts of speech (e.g.,「noun + noun」,「symbol + noun」,「auxiliary + adjective」, etc.). In the analysis by LOOCV (Leave-One-Out-Cross-Validation) for texts on another 100 blogs, the random forest model with 13 stylometric features showed the accuracy 80.0%: 81.3% for the "20s to 40s" age group and 79.4% for the "50s and 60s" age group in the rate of precision. Furthermore, the results of the support vector machines showed the accuracy 81.0%. The rates of precision were 78.4% for the "20s to 40s" age group and 82.5% for the "50s and 60s" age group. However, there was not statistical significant difference of the accuracy between both classifiers, this study displayed the possibility for the practical use of offender profiling.

Key words: offender profiling, age group estimation, stylometrics, random forests, support vector machines

キーワード：犯罪者プロファイリング，年齢層推定，計量文体学，ランダムフォレスト，サポートベクターマシン

機械学習を用いた著者の年齢層推定 —犯罪者プロファイリング実現に向けて—

財津亘，金明哲

1. はじめに

本邦犯罪捜査では，犯罪者プロファイリングと呼ばれる捜査支援手法を活用することで，犯人の性別，年齢層，職業などを推定し，犯人の早期検挙ならびに犯罪捜査の効率化を推進している [1]．犯罪者プロファイリングは，犯罪現場に関する情報などを基に実施するものであるが，印字文書や電子メール，電子掲示板への書き込みなど犯罪現場といえるものが存在しない事件もある．従来の犯罪者プロファイリング手法では，このような事件に対応ができなかった．このような状況から，財津・金 [2]は，ランダムフォレスト(RF: Random Forest)とサポートベクターマシン(SVM: Support Vector Machine)を用いて，文章情報から「性別」の推定を試みている．財津・金 [2]で

* Toyama Prefectural Police Headquarters, Toyama E-mail: wataru0112csi@yahoo.co.jp
** Faculty of Culture and Information Science, Doshisha University, Kyoto E-mail:mjin@mail.doshisha.ac.jp

は, まず 100 名のブログ(Blog)における文章情報に基づき，性別推定に有効と考えられる文体的特徴を検討している．続いて，別の 100 名のブログの文章情報を用いて，性別推定に有効と考えられた文体的特徴を実装した RF および SVM によって性別の推定精度の検証を行っている．その検証結果によると，最高で 86.0%の適合率が得られたとされる(適合率については後述する)．本研究は，著者の「性別」を対象とした財津・金 [2]に続き，著者の「年齢層」を対象として検証することを目的とした．

計量的文体分析によって著者の特徴を推定する手法は，諸外国では主に「著者プロファイリング (authorship profiling や author profiling) [3,4]」，我が国では「テキストプロファイリング [5]」などと呼称されている．著者の特徴推定研究を概観すると，情報工学系やコンピュータサイエンス系の研究者が携わっていることが多く，年齢層推定研究に関しても同様である．先行研究によると，ブログやエッセイコーパスを対象に，サポートベクターマシンや決定木，ブースティングといった分類器が使用されている [4,6-8]．

まず, 英語圏である諸外国の先行研究を概観する．Schler, Koppel, Argamon, Pennebaker[6]は，71,493 名のブログを対象に，性別ならびに年齢層別で使用頻度が異なる単語を検討し，推定精度の検証も合わせて試みている．報告によると，10 代では「maths」や「homework」，「bored」といった語が多く，20 代では「apartment」，「student」，「college」などが，30 代では「marriage」，「son」などの語が増加する傾向にあるという．このような単独で意味を有する語は内容語と呼ばれる．他方，単独では語彙的意味を持たないものの，文法機能を有する語を機能語という．この研究では, さらに内容語や機能語, ブログに特有な語, 品詞に着目し，MCRW(Multi Class Real Winnow)という分類器を用いて，年齢層(10 代[13-17 歳]・20 代[23-27 歳]・30 代中心[33-42 歳])の推定を試みており，全体の正解率は 76.2%であったとされる．なお，この種の推定精度に関する研究では，その評価指標として，正解率，適合率，再現率，F 値がある [9]．正解率とは，分類器によって算出されたすべての結果の中で，正しく推定された割合のことである．適合率とは, 分類器によって算出された結果が正しい割合, たとえば「20 代」と推定してその中で正しかった割合を意味する．一方，再現率とは，あるグループ内で分類器が正しい結果を算出した割合で，「20 代」のサンプル中，正しく「20 代」と分類した割合のことである．F 値とは，適合率と再現率の調和平均によって算出される値で, 両者を折衷した評価指標とされる．犯罪者プロファイリングでは，捜査員に提示する推定結果がどの程度「当たる」かが重んじられる．たとえば，分析者が「犯人は男性の可能性が高い」と提言する場合に，その分析結果を受け取る捜査員はこの分析結果がどの程度的中するものかに興味がある．このことから，正解率や再現率に比べると，適合率が重要といえる．そこで，前述の Schler et al.[6]の分析結果を基に適合率を計算したところ, 10 代で 85.4%, 20 代で 78.2%, 30 代中心で 45.1%であった. Argamon & Koppel[3]も，ブログ(19,320 名)を対象に，著者の性別ならびに年齢層の推定を行っている．検証の結果によると，10 代では「school」や「bored」などの内容語が，20 代では「apartment」や「office」，「work」，30 代の場合は「wife」や「husband」，「children」といった内容語が頻出するという．この結果は，Schler et al.[6]と類似しており，人々の興味が年齢層によって異なることを示している．また，機能語に関しては，10 代は「im」や「so」，「cant」などが，20 代や 30 代では「of」や「in」などの前置詞が相対的に多いとされる．この３分割区分(10 代・20 代・30 代)における正解率は，内容語のみを用いた場合で 75.5%，機能語のみを用いた場合で 66.9%，内容語と機能語を用いた場合で 77.7%であった(適合率は不明)．また, Santosh, Bansal, Shekhar, Varma[4]は，内容語や機能語に着目し，決定木を用いて年齢層の推定精度を検証している．報告によると，正解率は英語で 64.08%，スペイン語で 64.30%であったという(適合率は不明)．これら諸外国の先行研究ではすべて，内容語に着目しているものの，内容語は，語彙的な意味を有する語であり，話題内容もしくはジャンルに依存する．犯罪が関与する文章は，ブログと話題内容やジャンルが異なるため，本論文では，内容語ではなく，機能語を用いた．また，諸外国の先行研究では 10 代から 30 代を

対象としている研究が多いが，犯罪を敢行する者はこの年齢層に限らない．なお，諸外国の先行研究を概観すると，推定精度は，正解率でおおむね６割から７割後半程度といえよう．

翻って，本邦の年齢層推定研究は，先行研究が少ないのが現状といえる．Izumi, Miura, Shioya [10]は，Schler, et al.[6]にならい，10 代[13-17 歳]・20 代[23-27 歳]・30 代中心[33-42 歳], その他の年齢層に分類し，ナイーブベイズによる推定を実施し，71%の F 値が得られたという．また，萩野谷 [5,11]は，犯罪者プロファイリングへの応用を視野に入れた年齢層推定研究を行っている. 萩野谷 [11]は, 助詞の n-gram に着目し，20 代から 60 代を対象に，２群に分ける境界を変化させ，その２群について判別分析による検証を行った．その結果によると，30 歳を境界として，20 代と 30 代以上の２群について判別した場合で最も精度が高い 78.2%という成績が得られたという．ただし，この成績は適合率や再現率ではなく, 正解率である. 極端な話，すべて「30 代以上」と推定したとしても，サンプルの多くが 30 代以上であったため, 正解率は高くなったが，正しく「20 代」と推定している割合は不明であり，実質的に意味のある検証結果ではないといえる．また，萩野谷 [5]では，助詞の n-gram に加えて，読点前の文字を使用し，判別分析に比べて成績が良いとされている RF を用いている．その報告では，20 代から 60 代の５分割区分における推定結果が表になっている．それを基に適合率を算出してみると, 20 代で 66.0%, 30 代で 49.5%, 40 代で 22.5%, 50 代で 38.0%, 60 代で 72.0%であった. 最低で 22.5%, 最高で 72%の適合率となると，実務上有用なレベルとはいえないであろう．萩野谷 [5]は，文体的特徴として，助詞の unigram と読点前の文字に着目しているが，たとえば読点前の文字の中でも，年齢層推定に有効な特徴は一部のみであることが推察される．推定精度を向上させるには，年齢層別で特徴量が異なる文体的特徴を抽出することで，年齢層推定に有効な文体的特徴を用いるべきであろう．以上のとおり，本邦の研究においては最高で７割程度の成績であり，年齢層推定の難しさを示している．その他の年齢に関連する研究として，泉・三浦 [8]は，年齢層ではなく，実際の年齢に許容誤差を設定することで検証を行っている．また，岩崎・佐藤・駒谷 [7]は年齢層の推定ではなく，著者の生まれた年代の推定を行っている．

そもそも，著者の文体的特徴というものは，加齢にともない変化するものなのであろうか．この問いに対する一つの例として日蓮遺文の助詞の使用率に着目した村上 [12]の研究が挙げられる. 執筆年を横軸, 助詞の使用率を縦軸に図を作成すると，加齢にともない助詞の使用率が減少傾向にあり，ある時期を境に分散が大きくなるといった傾向がみられた．分析の結果によると，加齢にともなって減少傾向にあった助詞の使用率は，ある時期を境にその傾向や分散が変化したと結論付けている．１事例ではあるが，著者の文体的特徴が変化する例であり，またこのような加齢にともなって変化する文体的特徴が年齢層の推定に有効であるといえよう．

犯罪者プロファイリングの目的は，犯罪捜査の効率化にあり，そのために犯罪捜査の範囲を絞り込むことである．通常は，捜査員に対してピンポイントに年齢を提言することはなく，「犯人は 20 代から 30 代の可能性が高い」など年齢幅として表現する. また, 犯人像推定の結果は，それ以降の犯罪捜査にも影響を与える可能性があることから，当然高い推定精度が求められる．先行研究 [5]のような５分割区分(20 代から 60 代)では，適合率で最低 22.5%，最高 72.0%とかなり低かった．このことから，本研究では，２分割区分に設定し，適合率においておよそ 80%程度の推定精度が得られるモデルの構築を目指した．そこで，まず加齢にともない特徴量が増加あるいは減少する文体的特徴を探索的に検討し，続いて別のブログサンプルを使用し, それらの文体的特徴を用いて, 80%程度の適合率が得られるモデルの構築を目指した．なお，機械学習における分類器は複数存在するが，前述のとおり，従来の著者特徴の推定研究では SVM が多用されている．また，先行研究 [5]では判別分析に比べて, RF の成績が比較的高かったとされる. このことから，本研究では両機械学習法を用いて比較検討を行った．

2. 方法

2.1 サンプル

著者の性別推定を題材とした財津・金 [2]と同一のサンプルを使用することとし，インターネットサイト「Yahoo!ブログ(https://blogs.yahoo.co.jp/)」と「にほんブログ村(http://diary.blogmura.com/)」のブログから抽出した．

抽出方法は，それぞれのブログサイトにおいて，２つの性別(男性，女性)×５つの年齢層(20 代から 60 代)の 10 グループで，各 10 名のブログ(計 100 名)を無作為に選定し，さらにもう一つのブログサイトから，同様の方法でブログ(計 100 名)を抽出した(全サンプルで計 200 名分のブログ)．

2.2 年齢層で特徴量が異なる文体的特徴の探索

年齢層の推定に有効な文体的特徴を探索するために，「Yahoo!ブログ」における 100 名のサンプルを用いて，加齢にともない特徴量が増加あるいは減少する傾向があるなど，年齢層によって特徴量が異なる文体的特徴を検討した．テキストの作成および着目した文体的特徴については，財津・金 [2]と同様である．

2.2.1 テキストの作成

文字数の影響を統制するために，著者１名につき 1,000 文字以降の最初の末尾までの文章を用いて，テキスト形式に変換した電子データであるテキストを作成し，総数 100 テキストを分析に用いた．

2.2.2 特徴量の算出

各テキスト内の，以下の文体的特徴に着目し，「度数」と「相対度数」を算出した：漢字，平仮名，片仮名，ローマ字，数字，文字の unigram，bigram，trigram，読点の打ち方(読点前の文字や単語)，文の長さ(文字数)，品詞の unigram，bigram，trigram，単語の unigram，bigram，trigram，単語の長さ(文字数)，句点の打ち方(句点前の文字や単語)．「度数」と度数から変換する「相対度数」，品詞情報の第１階層(例，助詞)から第３階層(例，助詞－格助詞－連語)のそれぞれについて検討した点は，財津・金 [2]と同様である．

2.2.3 各年齢層の特徴量の比較

それぞれの文体的特徴の度数もしくは相対度数を算出した後に，年齢層(20 代から 60 代)グループごとで特徴量の変動を把握するとともに，各特徴量に関して，χ^2 検定(5%水準)を実施することで，年齢層別における特徴量の相違について検討した．

2.3 年齢層推定にかかる精度の検証

続いて，「にほんブログ村」における 100 名のサンプルを基に，年齢層を推定する上で有効と考えられた文体的特徴を機械学習に学習させ，その推定精度を交差確認法によって検証した．

萩野谷 [5]が示したとおり，年齢層を５分割区分した場合には，一部の年齢層で推定精度がかなり低下することが予想された．そこで，ある程度の推定精度を保つために本研究では，５つの年齢層を２分割し，次の分割区分で検証を行った．

- 分割区分①「20 代 vs. 30 代から 60 代」
- 分割区分②「20 代 30 代 vs. 40 代から 60 代」
- 分割区分③「20 代から 40 代 vs. 50 代 60 代」
- 分割区分④「20 代から 50 代 vs. 60 代」．

2.4 機械学習法

本論文では，以下の機械学習法を用いた．

2.4.1 ランダムフォレスト

Breiman[13]が提案した集団学習法の一種で，学習用データから学習された多数の決定木の結果を組み合せるまたは統合して成績を向上させるといった機械学習法である．本研究では，財津・金 [2]と同じく，データセットを基に構築される 30,000 本の決定木による多数決により，相対的に良い決定木のモデル構築を行った．また，特徴量の数に関しても，最適な設定の探索を行ったところ，すべての分割区分において，特徴量の数が２で最も高い推定精度であった．

本研究では，事前に年齢層推定に有効と考えられる文体的特徴の抽出を行うが，それらの文体的特徴における識別力を，「にほんブログ村」のサンプルを用いた RF によって算出される指標で評価した．識別力の指標には，Mean Decrease Accuracy (MDA)を用いた．MDA は，ある文体的特徴を除いた場合に低下する推定精度の程度を示した指標で，値が大きいほど識別力があることを意味する．

2.4.2 サポートベクターマシン

Vapnik [14]が開発した教師あり学習法の一種である．学習用データを基に，無限に存在する超平面の中から，各データ点との距離が最大となるマージン，つまりは最もグループ分けの良い超平面を探索する機械学習法である．

SVM のパラメータは，RF に比べると数が多く，その設定により成績が大幅に異なる場合がある．このことから，財津・金 [2]にならい，最適なパラメータの値を探索すべく，カーネルに RBF(radial basis function)を用いた非線形 SVM において，C(誤分類の許容レベル)と γ(境界線の複雑さ)に関するグリッドサーチを実施した．加えて，正規化の有無や線形カーネルについても検討した．

2.5 交差確認法（１個抜き）による検証

検証は，１個抜き交差確認法 (LOOCV: leave-one-out cross-validation)によって実施した．まず，「にほんブログ村」におけるすべてのサンプル(100 名)から１サンプル(１名)のみを抽出し，残りのサンプル(99 名)を学習用データとして学習させる．続いて，抽出した１サンプルを予測用データとして用いて年齢層を推定する．これを全 100 サンプルに対して，順次繰り返して行った．

以上の LOOCV による検証結果に基づき，両分類器ごとに正解率，適合率，再現率，F 値を算出した．F 値については，(２×適合率×再現率)/(適合率+再現率)にて F_1 値を算出した．

2.6 ソフトウェア

形態素解析が可能な「茶筌(ChaSen)[15]」や機械学習法による分析が可能な「R(ver.3.2.3) [16]」が実装されている多言語テキストマイニングツール「MTMineR (ver.5.3) [17]」を使用した．

3. 結果

3.1 年齢層推定に有効な文体的特徴

Table 1 に，年齢層のグループ別における文体的特徴の度数および検定の結果を示す．これらの文体的特徴は，①名詞，②読点前の「は(係助詞)」，③「です(助動詞)」+「けど(接続助詞)」，④「ずっと(副詞)」，⑤品詞の bigram(「名詞+名詞」や「副詞+副詞」,「助動詞+形容詞」など)の使用頻度として大分類できよう．Table 1 で示したとおり，加齢にともない，特徴量が増加する傾向があるもの(たとえば，名詞(一般)や読点前の「は(係助詞)」の使用頻度など)と特徴量が減少する傾向(たとえば，名詞(接尾－助動詞語幹)や「ずっと(副詞)」など)の文体的特徴がみられた．χ^2 検定を実施した結果によると，ほとんどの文体的特徴が５％水準で有意差がみられた．ただし，「です(助動詞)＋けど(接続助詞)」および品詞の bigram「副詞+副詞」については，５％水準において有意差はみられなかった．しかしながら，p 値が有意水準と同程度であったことから，続く推定精度の検証時には，これらの文体的特徴もモデルに組み込んだ．

3.2 年齢層推定の精度検証

次に，Table 1 における 13 の文体的特徴を用いて，前述した年齢層の分割区分ごとに，LOOCV による年齢層推定の精度検証を行い，正解率，適合率，再現率，F_1 値を算出した(Table 2(a)～(d)).

Table 2 の正解率をみると，両機械学習法ともに分割区分③において，80%程度の最も高い推定精度が得られた．適合率や再現率，F_1 値についても，おおむね分割区分③が，その他の分割区分に比べて高い精度を得たものといえよう．正解率を機械学習法別でみると，分割区分①と④で RF が SVM に比べて推定精度が多少高かったのに対して，分割区分②と③においては SVM の方が RF に比べて推定精度が高い傾向にあった．分割区分②については，両機械学習法の推定精度(正解率)の差が６%ほどあったものの，その他の区分については，ほとんど差がないといえる．犯罪者プロファイリングでは前述のとおり適合率が重要であるが，着目すべき点は，RF における分割区分①の「20 代」において適合率が 33.3%，分割区分④の「60 代」の適合率にいたっては０%であった．RF ほどではないものの，SVM の適合率も，分割区分①の「20 代」で 16.7%，分割区分④の「60 代」で 33.3%と実務への応用は困難と言わざるを得ない．一方で，分割区分③については，RF・SVM ともに，８割程度の適合率が得られ，実務上の分析における有用性を示したといえよう．

Table 1. Frequencies classified by age group and results of statistical analysis.

Writing Style	Frequencies classified by age group					χ^2 test	MDA[*1]
	20s	30s	40s	50s	60s		
Noun(General)	1482	1625	1524	1684	1805	$p < .01$	4.9
Noun(Suffix.measure)	58	110	73	93	143	$p < .01$	3.7
Noun(Suffix.auxiliary)	9	16	8	5	3	$p < .05$	3.8
「は(Binding particle)」 just before comma	35	36	38	65	75	$p < .01$	4.5
「です(Auxiliary)」 ＋「けど(Conjunction)」	5	4	0	2	0	$p = .05$	1.5
「ずっと(Adverb)」	7	10	3	1	0	$p < .01$	0.5
Noun＋Noun	599	730	663	748	879	$p < .01$	4.8
Noun(Numeral)＋ Noun(Suffix.measure)	46	95	60	75	132	$p < .01$	3.9
Symbol＋Noun	633	777	804	847	960	$p < .01$	4.8
Particle(Pronominal)＋ Noun(General)	207	217	190	244	294	$p < .01$	4.9
Auxiliary＋Symbol	367	337	408	387	494	$p < .01$	4.8
Auxiliary＋Adjective	22	7	13	13	10	$p < .05$	1.4
Adverb＋Adverb	12	6	11	16	4	$p = .05$	2.8

[*1] MDA indicates importance of variable

13 の文体的特徴における MDA を概観すると,「名詞(一般)」や「助詞(連体化)＋名詞(一般)」の使用頻度において最も識別力を有することが分かる(Table1). 他方,「ずっと(副詞)」の使用頻度は，MDA が最も低かった．また，Table 1 の度数と MDA をみると，全体的に使用頻度の低い文体的特徴で MDA が低い傾向がうかがえる．このことは，年齢層に関する識別力が文章内の使用頻度に影響し，使用頻度の高い文体的特徴の方が年齢層の判別に有効であることを示している.

4. 考察

本研究では，まず年齢層別で特徴量に違いがみられる文体的特徴を探索したところ，Table 1 の 13 の文体的特徴が選出された. 萩野谷 [5]は, 助詞の n-gram と読点前の文字に着目していたが，実際に年齢層推定に有効と考えられる文体的特徴は, 助詞の unigram の中でも「ずっと(副詞)」のみであり，読点前の語については，「は(係助詞)」の使用頻度のみが有用と考えられた．その他の文体的特徴については，先行研究で指摘されていないことに加えて，文章内の品詞

の割合などであることから，話題内容による影響はほとんどなく，汎用性の高い文体的特徴であると考えられる．以上を踏まえた上で，分割区分別で推定精度を検証したところ，「20 代から 40 代」と「50 代・60 代」に２分割した場合の適合率が，RF・SVM ともに，80%程度であったことから，捜査支援手法として実務上応用が可能なものと思われる．

ただし，本研究で使用したサンプルは，ブログから抽出したものであり，自己の文章表現を意図的に変える必要があまりない，つまりは自然状態で記載されたものと考えられる．言い換えれば，本研究の結果は，自然状態で記載された場合の結果に限られるかもしれない．犯罪が関与する文書などを扱う場合には，そのような文章における特有の「作為性」と

Table 2. Accuracy, precision, recall, and F value in each age group.

a) 20s vs. 30-60s

Age Categories	R F			S V M		
	Precision	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
20s	1/3 (33.0%)	1/20 (5.0%)	8.7%	1/6 (16.7%)	1/20 (5.0%)	7.7%
30－60s	78/97 (80.4%)	78/80 (97.5%)	88.1%	75/94 (79.8%)	75/80 (93.8%)	86.2%
Accuracy	79/100 (79.0%)			76/100 (76.0%)		

b) 20-30s vs. 40-60s

Age Categories	R F			S V M		
	Precision	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
20・30s	18/32 (56.3%)	18/40 (45.0%)	50.0%	24/36 (66.6%)	24/40 (60.0%)	63.1%
40－60s	46/68 (67.6%)	46/60 (76.6%)	71.8%	48/64 (75.0%)	48/60 (80.0%)	77.4%
Accuracy	66/100 (66.0%)			72/100 (72.0%)		

c) 20-40s vs. 50-60s

Age Categories	R F			S V M		
	Precision	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
20－40s	54/68 (79.4%)	54/60 (90.0%)	84.4%	52/63 (82.5%)	52/60 (86.7%)	84.5%
50・60s	26/32 (81.3%)	26/40 (65.0%)	72.2%	29/37 (78.4%)	29/40 (72.5%)	75.3%
Accuracy	80/100 (80.0%)			81/100 (81.0%)		

d) 20-50s vs. 60s

Age Categories	R F			S V M		
	Precision	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
20－50s	77/97 (79.4%)	77/80 (96.3%)	87.0%	72/88 (81.8%)	72/80 (90.0%)	85.7%
60s	0/3 (0%)	0/20 (0%)	0%	4/12 (33.3%)	4/20 (20.0%)	25.0%
Accuracy	77/100 (77.0%)			76/100 (76.0%)		

いった問題が挙げられる. 筆跡鑑定の例を挙げると, 意図的に他人の筆跡を模倣する模倣筆跡や自己の筆跡を隠蔽する目的で意図的に変える韜晦筆跡といった作為筆跡が存在する[18]. 文章の表現方法も同様で, 犯罪行為を行う意図を持つ著者が何らかの文章を記載する場合, 日常の自然状態で記載している表現方法を意図的に変える可能性がある. たとえば, 40 代の女性が, 10 代の男子学生にストーカー行為をしていて, 10 代の女子学生を装って手紙を書くといったことも考えられよう. ただし, 吉田[18]によると, 模倣筆跡の場合は, 字画の長さや文字全体の形が手本の筆跡に似ていても, 入筆部や転折部, 終筆部に筆者本人の個性が表出する傾向があるという. また, 韜晦筆跡の場合は, 文字全体の外形や字画の長短などが変えられる傾向があるが, 入筆方向や筆順などは筆者本人の個性が検出されるという. これは,「潜在的筆跡個性」という筆者自身が気付いていない筆跡の個性が表出するために筆者本人の個性を検出することができるとされている. 同様に, 村上[19]も, 無意識に書く所に, 個人の文章の特徴が現れやすいとされており, 文章表現にも通じると考えられる. たとえば, 本研究における名詞や機能語の使用頻度といった文体的特徴は, 無意識に書くことが推察される. このことから, いわゆる「作為性」による影響は, さほど無いものと考えられるが, 年齢層を偽る場合には, どのような文体的特徴が原文から変化するものなのかを検討する必要はあろう. 以上から, 今後の課題として, 年齢層を意図的に偽って文を記載してもらうなどの実証実験を実施することが求められる.

以上の結果は, ブログのサンプルを分析することで得られたものであるが, 本研究で用いたブログに登録されている性別や年齢がどの程度正確かは不明である. 実際の年齢を偽っている可能性もありえよう. したがって, 今後は実際に幅広い年齢層の参加者を募って, 文章を書かせる課題を行うなど, 年齢を確実に把握できうるサンプルで検討するといったことも必要と考えられる.

先行研究[5]では, 年齢層を５分割した場合に, かなり低い推定精度が示されていた. このことから, 本研究では, 推定精度の保持といった観点から, 年齢層を２分割することで推定精度の検証を行った. ただし, 犯人の絞り込みといった目的から考えると, ３分割ないし４分割など, より犯人の絞り込みを行うことが可能な分割方法が求められる. このことから, 今後は推定精度の向上とともに, ある程度の推定精度を保持できうる分割区分についても引き続き検討する必要がある.

「性別」と異なり,「年齢層」は時間の経過とともに変化するものでもある. 20 年前の 30 代(現在の 50 代)と現在の 30 代では当然文体的特徴が異なることが容易に予想できる. したがって, 年齢層推定に関する研究は, 随時テキストデータを更新し, 研究を続けていかなければならないであろう.

性別推定と同様に, 著者の知る限り, 我が国の犯罪捜査において, 文章情報を基に年齢層を推定した犯罪者プロファイリング事例は存在しない. しかしながら, 実際の犯罪捜査場面では, 被疑者が不詳である電子掲示板の書き込みや電子メール, また印字文書による犯罪は後を絶たない. また, 今後もこれらの媒体を使用した犯罪は増加しても減少することはないであろう. このような犯罪捜査情勢であることからも, 本手法が犯罪者プロファイリングの新たな一手法として確立されることが期待される. 実務への応用に向けては, 推定精度の向上を課題とする他に, 作為性を有する文章の影響や犯罪が関与した文章を検証する必要があろう.

参考文献

1) 警察庁, “平成 27 年版警察白書”, (2016).
2) 財津亘, 金明哲, “ランダムフォレストによる著者の性別推定－犯罪者プロファイリング実現に向けた検討－”, 情報知識学会誌, **27**[3], 261-274 (2017).
3) S. Argamon, M. Koppel, “A Systemic Functional Approach to Automated Authorship Analysis”, *Journal of Law and Policy*, **21**[2], 299-315 (2013).
4) K. Santosh, R. Bansal, M. Shekhar, V. Varma, “Author Profiling: Predicting Age and Gender from Blogs”, *Notebook for PAN at CLEF*, 119-124 (2013).
5) 萩野谷俊平, “テキストプロファイリングによる犯人像推定の検討”, 犯罪心理学研究, **48**[特別号], 134-135 (2010).

6) J. Schler, M. Koppel, S. Argamon, J. Pennebaker, “Effects of Age and Gender on Blogging”, *Proc. of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report*, **6**, 199-205 (2006).
7) 岩崎裕也，佐藤理史，駒谷和範，“エッセイコーパスを用いた著者の生年の推定”，言語処理学会第 19 回年次大会発表論文集，652-655 (2013).
8) 泉雅貴，三浦孝夫，“ブースティングに基づく Blog 著者年齢推定”，*DEIM Forum* (2009).
9) 金明哲，テキストデータの統計科学入門，(岩波書店，東京，2009).
10) M. Izumi, T. Miura, I. Shioya, “Entropy-Based Age Estimation of Blog Authors”, *IEEE Annual International Computer Software and Applications Conference*, 795-800 (2008).
11) 萩野谷俊平，“文体による筆者の識別と特性推定”，犯罪心理学研究，**47**[特別号]，114-115 (2009).
12) 村上征勝，“計量文献学―文献の新たな研究法―”，村上征勝，金明哲，土山玄，上阪綾香，計量文献学の射程，(勉誠出版，東京，2016)，pp.7‐55.
13) L. Breiman, “Random Forests”, *Machine Learning*, **45**[1], 5-32 (2001).
14) V. Vapnik, *The Nature of Statistical Learning Theory*, (Springer, NewYork, 1999).
15) 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).
http://chasen-legacy.osdn.jp/（2016 年 1 月 15 日参照）
16) R Development Core Team, “R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing”(2015).
https://cran.r-project.org/（2016 年 1 月 15 日参照）
17) 石田基広，金明哲 編著，コーパスとテキストマイニング，(共立出版，東京，2012).
18) 吉田公一，ポイント解説 筆跡・印章鑑定の実務，(東京法令出版，東京，2004).
19) 村上征勝，シェークスピアは誰ですか？―計量文献学の世界―，(文藝春秋，東京，2004).