Vol.2020-IFAT-138 No.3
Vol.2020-DC-116 No.3
2020/3/27

文単位での著者識別に基づく公的文書の改ざん検知手法の検討

松林由佑[1]　二又航介[2]　猪俣敦夫[3]　井上博之[4]　衛藤将史[5]

概要：今や生活の一部とも言えるインターネット上には実に様々な文書が掲載され，その中には著作権を侵害した形態での掲示や改ざんして掲載する等非常に深刻な問題を引き起こしているものがある．いわゆる「モリカケ問題」をはじめとして，公的な文書の書き換えが問題となることがあるが，それらを自動的に検知・特定することは，元の文章が存在しない限り非常に難しい．そこで本研究では，書き換えられた部分と元の部分の書き手が異なるという視点から，著者識別のみを対象として研究されてきた著者識別の手法は存在するものの公的文書に対しては未だ不十分と言える．そこ技術を利用して改ざんを検知する手法について検討を行った．これまで文学的作品など私的文章ので，本項では公的文章の改ざん検知に適用する方法を検討し，その手法の有用性を評価するために実施した実験結果を報告する．

キーワード：著者識別，改ざん検知，計量文体学

1. はじめに

書籍や論文をはじめとして著作者に権利のある文書等の著作物に対する盗用や改ざん等の行為は今もなお大きな問題となっている．特に，インターネットでは，実に様々な文書が掲載されており，その中には著作権を侵害した形態での掲示や改ざんして掲載するなど非常に深刻な問題を引き起こしているものがある．このため，文書そのものの原点を特定するという意義において，例えば文書の著者を識別する手法等を確立することはセキュリティの観点においても喫緊の課題であると言える．

本研究では，書き換えられた部分と元の部分の書き手が違うという考えから，著者識別の技術を利用して改ざんを検知する手法について検討を行った．これまで限られたデータセットでのみ行われてきた著者識別の手法を，公的文書の改ざん検知に適用する方法を検討し，その手法の有用性について評価するために実施した実験結果を報告する．

2. 既存研究と課題

著者識別については，学術論文等の著作物に記載されている著者や所属情報が不十分な場合において，著作物を管理するデータベース等の情報を用いて著書を特定する際に用いられることが多く, Strotmann らによって学術文献における著者識別問題を指摘している[1]．しかしながら，データベースが保有する情報やそもそも情報が欠落している状態においては著者識別を見誤るリスクが伴う．このため，著者識別においては文書そのものに対しての検討が重要な要素であると考える．そこで本稿において検討する著者識別とは，ある著者の分からない文章に対して，その文章の著者を推定するタスクであると定義する．

著者識別に関しては様々な研究が行われている．例えば，三品・松田らは，小説やブログ記事に対して，読点前の文字や単語の長さ，助詞の分布などを用いて[2]，また，金らは学生作文と文学的作品について，ランダムフォレストや SVM など 6 手法を組み合わせて[3]，出力の多数決をとることで最終的な出力を決定するという統合的分類アルゴリズムによって，それぞれ高い精度で著者を同定することに成功している．

しかしながら，例えば公的に発行される文書や官報などのような公的文書に対する著者識別の必要性についての議論は不十分な状況であり，またそのような公的文書に対する研究は行われてはいない．このように既存手法においては, 著者が有する書き手の特徴に着目していることにより，より著者を識別するための情報が現れやすいとされる私的文章を対象にしている．一方，公的文書等の文章においては，文化庁により定められている「公用文の書き方資料集」[4]に基づいた書き方のガイドラインに従うことにより，より著者個人の文体的特徴が出にくくなるという点が挙げられる等において私的文章との違いがみられる．

また，識別する文章の単位に関して，PAN style change detection[5]等に見られるように，文章全体やパラグラフ単位での著者識別は盛んに研究されているが，文単位での著者識別の事例は少ないのが現状である．

3. 公的文書における改ざん検知

3.1 対象とする文書

公的文書を対象に文単位での識別を行うことは改ざんの検知につながるため，本手法については十分な需要があると考えられる．はじめに，提案手法を説明するにあたり，表 1 に改ざん文章の例を示す．本文章は，財務省が報告した森友学園案件に関わる決裁文書の改ざんに関する調査[6]

1 滝高等学校
2 奈良先端科学技術大学院大学
3 大阪大学
4 広島市立大学
5 情報通信研究機構

表 1　改ざんされた公的文書の例

改ざん前の文	改ざん後の文
地質調査会社に当該ボーリング調査結果をもとに本地の地理に意見を求めたところ，特別に軟弱であるとは思えないとしたうえで，通常と比較して軟弱かどうかという問題は，通常地盤の定義が困難であるため，回答は難しいという見解であった.	ボーリング調査結果について，専門家に確認するとともに，不動産鑑定評価を依頼した不動産鑑定士に意見を聴取したところ，新たな価格形成要因であり，賃料に影響するとの見解があり，価格調査により，鑑定評価を見直すことにした.

の報告書の一部である．本研究では，文学作品やブログ記事に用いられた方式を公的文書に適用することで，上記のこのような部分的な改ざんに対応し，文単位での文書の著者識別における精度を検証することを目的とする．

3.2 データセット

本研究では，公的文書の代表例として表 2 に示す 10 省庁の白書等を対象に実験を行う．選定では語尾が「だ・である調」であることに留意した．なお，これらの白書の中には，特集として，一部違う著者の書いた文章が混ざっていたが，それらは手動で削除した．

また，比較対象として，表 3 に示す文学作品についても同様の方法で識別を行う．

3.3 著者識別に対する特徴量

著者の識別を行う上でどのような特徴量を用いるかは多くの研究が行われているが，本研究では財津・金[7]等にならい，代表的な特徴量として以下の 3 特徴量を採用することとする．また，それらを統合した特徴量データを評価に用いた．

「品詞と Bi-gram 出現頻度」，「機能語と Bi-gram 出現頻度」を調べるにあたり，文章を単語単位で分割する必要があるが，それには形態素解析ソフト MeCab[8]を用いた．

・品詞と Bi-gram 出現頻度

品詞とは，単語をその機能や形態などで分類したものである．著者識別の上で用いられる品詞として，「名詞」「動詞」のように，品詞単体のものと「名詞-一般」「助詞-係助詞」のようにタグが付いたものの 2 種類があるが，本研究では両者を統合したものを用いた．実際に，カットオフ値を設定し, 全体の中で 1000 回以上出現する要素を用いたところ，ベクトルの長さは 434 であった．

表 2　対象とした 10 省庁の文書

表題	著者/官庁	単語数	文数
年次経済財政報告	内閣府	45312	1155
厚生労働白書	厚生労働省	39581	992
通商白書	経済産業省	11094	264
地方財政白書	総務省	11734	297
外交青書	外務省	40185	988
公務員白書	人事院	19108	436
警察白書	警察庁	23080	429
会年次報告	公正取引委員会	94942	1587
中小企業白書	中小企業庁	32457	1093

表 3　比較対象の文学作品

表題	著者	単語数	文数
地獄変	芥川龍之介	8216	214
婦系図	泉鏡花	28984	1068
無名作家の日記	菊池寛	10229	441
山椒大夫	森鴎外	6176	333
三四郎	夏目漱石	55691	3346
桃の雫	島崎藤村	12876	573
断崖の錯覚	太宰治	56966	1560
青蛙堂鬼談	岡本綺堂	5860	294
怪星ガン	海野十三	52112	1670
風の又三郎	宮沢賢治	37744	1705

・機能語と Bi-gram 出現頻度

機能語とは，「は」「も」など，単体で意味をなさず，文内の他の単語との関係性を示す単語を指す．実際に，全体の中で 1000 回以上出現する要素を用いたところ, ベクトルの長さは 115 であった．

・文字と Bi-gram 出現頻度

全体の中で 2500 回以上出現する要素を用いたところ，ベクトルの長さは 63 であった．なお，文字 Bi-gram 出現頻度のカットオフ値を 2500 としたのは，「学-校」「障-害」など，著者の特徴ではなく内容に依存する要素を排除するためである．

3.4 N-gram の適用

N-gram とは, 任意の文字列や文書を連続した N 個の文字で分割するテキスト分割を用いた手段であり，その中でも隣接 2 要素について適用したものは Bi-gram と呼ばれる．著者識別に文字の N-gram を用いることの有効性は金らによって示されている[9]．例えば，「吾輩は猫である．」

Vol.2020-IFAT-138 No.3
Vol.2020-DC-116 No.3
2020/3/27

```
吾輩    ワガハイ        吾輩    名詞-代名詞-一般
は      ハ      は      助詞-係助詞
猫      ネコ    猫      名詞-一般
で      デ      だ      助動詞  特殊・ダ        連用形
ある    アル    ある    助動詞  五段・ラ行アル  基本形
```

図 1　MeCab による形態素解析

(品詞)Bi-gram：名詞-助詞，助詞-名詞，名詞-助動詞，助動詞-助動詞，助動詞-記号

(文字)Bi-gram：吾-輩，輩-は，は-猫，猫-で，で-あ，あ-る，る-.

図 2　Bi-gram の適用

という例文を品詞分解すると図 1 のようになるが，これに Bi-gram を適用した例を図 2 に示す.

本研究では，文章全体に対して上記 N-gram 変換を行い，各要素の文中における頻度を特徴量として使用した. また，カットオフ値を設定し，全体で出現回数が一定回数未満の要素は削除した. これは学習の効率化という目的のほかに，データセットの値がほとんど 0 になってしまうことから，正常動作しないことを防ぐためでもある.

3.5 ベクトル化

これらの分割された要素のうち，予め定めたものの文章中での相対頻度をベクトルに格納する．例文「吾輩は猫である．」を「品詞と Bi-gram 出現頻度」でベクトル化したものを表 4 に示す.

表 4　例文をベクトル化したときの頻度

名詞	40%
助詞	20%
（中略）	
名詞-一般	20%
名詞-代名詞	20%
（中略）	
名詞+助詞	25%
（中略）	
名詞-一般+名詞-代名詞	0%

3.6 分類器

本研究ではランダムフォレストとニューラルネットワークを分類器として用いた．次節以降にて，ランダムフォレストとニューラルネットワークの詳細について述べる.

3.6.1 ランダムフォレスト

ランダムフォレスト法とは，多数の決定木を使用したアンサンブルアルゴリズムの一つであり，著者識別における有効性は金[10]等によって示されている．本研究では精度の向上が目的ではないため，機械学習ライブラリ「scikit-learn」のデフォルト設定を用いて実験を行った.

3.6.2 ニューラルネットワーク

著者識別においてニューラルネットワークを用いた研究は多くないが，渡邊ら[11]によるとランダムフォレストを若干超える精度を達成したと報告されており，また，近年様々なタスクで応用されている手法であることから採用した．本研究では，割合 0.5 のドロップアウトを含めた 3 層のニューラルネットワークを用いた．最適化アルゴリズムには Adam を用い，学習率は tensorflow のデフォルト値を用いた.

3.7 データセットの分割とラベル付け

データセット分割の手順を以下に示す.

表 5　データのサンプリングの方法

	訓練データ	テストデータ
ラベル「著者 i」	「通商白書」から 8 割（212 文）をランダムに抽出	「通商白書」から残りの 2 割（52 文）を抽出
ラベル「著者 i でない」	残りの 7 省庁の文章全体のうち，212 文をランダムに抽出	「厚生労働白書」から 52 文をランダムに抽出

表 6　クロス表

	分類結果が著者 i	分類結果が著者 i 以外
著者 i の文章	A	B
著者 i 以外の文章	C	D

Vol.2020-IFAT-138 No.3
Vol.2020-DC-116 No.3
2020/3/27

(1) i, j（1≦i, j≦サンプル数）をランダムに選択
(2) 著者 i の文と著者 i 以外の文を「著者 i である」，「著者 i でない」でラベリング
(3) 著者 i の文章から，ランダムに 8 割サンプリングしたものと，著者 i, j 以外の文章をすべて訓練データとして適用
(4) 著者 i の文で，訓練データとして使用しない方の 2 割と著者 j についての分類を行い，精度を評価
(5) 1-4 手順の繰り返し

なお，著者 i であるか否かでラベル付けを行うと，2 つのラベルの個数に極端な偏りが発生したため，under-sampling（「著者 i でない」ラベルを他方のラベルと個数が等しくなるよう数を減らす手法）を適用した．これは，ニューラルネットワークにおいて訓練データのラベルに偏りがある場合，学習された分類器がラベルの個数の多い方に偏った判断をしてしまうからである．例えば，著者 i を「経済産業省（通商白書）」・著者 j を「厚生労働省（厚生労働白書）」にした場合の各データのサンプリングの方法を表 5 に示す．

3.8 評価指標

分類器による出力結果と実際のラベルを比較することで，表 6 に示すようなクロス表が作成できる．ここで，

再現率： $R = \dfrac{A}{A+C}$

適合率： $P = \dfrac{A}{A+B}$

と定義する．

再現率 **R** と適合率 P は一方が上がれば他方が下がるトレードオフの関係にあるため，**R** の平均値を $\hat{R}$，**P** の平均値を $\hat{P}$ として，その調和平均である

F 値： $F = \dfrac{2\hat{R}\hat{P}}{\hat{R}+\hat{P}}$

を，本研究での評価基準とした．ここで，F 値が 1 に近づくほど精度が高いと判断することとする．

4. 実験結果

ランダムフォレストによる実験結果を図 3 に，ニューラルネットワークによる実験結果を図 4 にそれぞれ示す．図 3 および図 4 における POS+bi-gram, FW+bi-gram, char+bi-gram, all はそれぞれ「品詞と Bi-gram 出現頻度」「機能語と Bi-gram 出現頻度」「文字と Bi-gram 出現頻度」を特徴量とした際の実験結果を示す．また，Formal Document

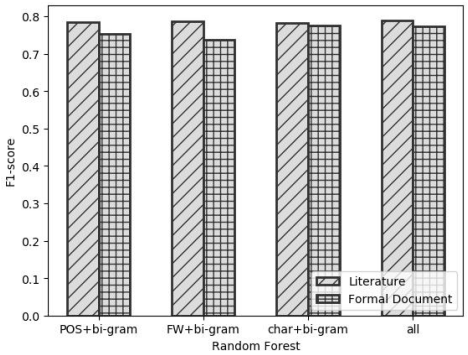

図 3　ランダムフォレストを用いたもの

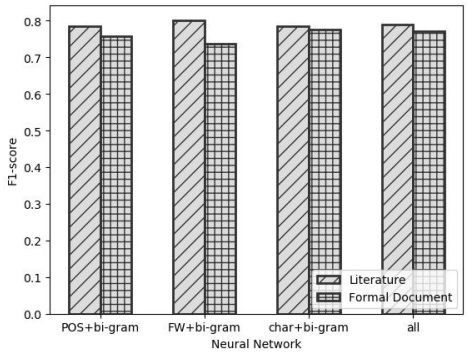

図 4　ニューラルネットワークを用いたもの

および Literture はそれぞれ表 2 に示す公的文書を対象とした識別精度，表 3 に示す文学作品を対象とした識別精度を示す．

図 3 に示すランダムフォレストによる実験結果および図 4 に示すニューラルネットワークによる実験結果では，特徴量選択による識別精度に有意な差がみられなかった．「品詞と Bi-gram 出現頻度」・「機能語と Bi-gram 出現頻度」を特徴量に用いた場合，公的文書は文学作品にやや劣る結果となったが，「文字と Bi-gram 出現頻度」を特徴量に用いた場合はほとんど差が見られなかった．これは，公的文書は文章の構造がある程度規格化されている一方で，そのような条件下においても文字単体の使用率には個人差が現れることを示している．

以下に誤判定の例を載せる．

「中南米地域の実質 GDP 成長率は、2018 年は前年比 1.0％と 2017 年の 1.3％からやや減速したが、緩やかな回復を続けている。」

これは，「機能語と Bi-gram 出現頻度」を特徴量として，実際には「著者 i でない」とラベルがついているが「著者 i である」と予測された文章である．調査すると，訓練データで「著者 i である」とラベルがつけられた以下の文が見つかった．

「公債費は、平成 18 年度以降低下の傾向にあったが、29 年度においては前年度と比べると 0.1 ポイント上昇の 12.9%となっている。」

この 2 文は使われている機能語が類似しており，特長量ベクトルのコサイン類似度が 0.96 と高い値を示した．このように，公的文書の短文においては，著者が異なる場合でも偶然似た文が現れる可能性があり，それが精度を引き下げている原因と考えられる．

5. まとめ

本稿では，書き換えられた部分と元の部分の書き手が違うという考えから，著者識別の技術を利用して改ざんを検知する手法について検討し，公的文書の改ざん検知に適用する方法を提案した．具体的には，著者識別の既存手法を公的な文章中の短文に適用する実証実験を実施し，その結果，文字の Bi-gram を特徴量として用いた場合にデータセットによらず安定した精度が出ることが確認された．しかしながら，分類器による精度の違いはほとんど見られなかった．今後の課題ではあるが，入力データの安定しない短文について検討する必要があるため，そのような文章に対しても適するより特化したベクトル化の手法を検討したい．

参考文献

[1] A. Strotmann, D. Zhao, and T. Bubela: Author name disambiguation for collaboration network analysis and visualization, Proc. American Society for Information Science and Technology, vol.46, no.1, pp.1-20, 2009.

[2] 三品光平,松田眞一: 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, vol.13, pp.35-46, 2013.

[3] 金明哲: 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, vol.41, no.1, pp.35-46, 2014.

[4] 文化庁: 公用文の書き方資料集, https://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/series/21/pdf/kokugo_series_021.pdf （参照 2020-02-15）.

[5] PAN: Style Change Detection 2020, https://pan.webis.de/clef20/pan20-web/style-change-detection.html（参照 2020-02-15）.

[6] 財務省: 決裁文書の書き換えの状況, https://www.mof.go.jp/public_relations/statement/other/201803B.pdf（参照 2020-02-17）.

[7] 財津亘, 金明哲: テキストマイニングを用いた著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―, 日本科学技術学会誌, vol.22, no.2, pp.91-108, 2017.

[8] 工藤拓，MeCab: Yet Another Part-of-Speech and Morphological Analyzer, https://taku910.github.io/mecab/

[9] 金明哲: 品詞のマルコフ遷移の情報を用いた書き手の同定, 日本行動計量学会 第 32 回全国大会講演論文集, pp.384-385, 2004.

[10] 金明哲, 村上征勝: ランダムフォレスト法による文章の書き手の同定, 統計数理, vol.55, no.2, pp.255-268, 2007.

[11] 渡邊翔, 松田眞一: 深層学習を用いた文章の書き手の同定, 南山大学紀要, vol.18, pp.1-13, 2018.

付録

実験に使用したデータセットの入手先を以下に示す.

(a) 内閣府：年次経済財政報告, https://www5.cao.go.jp/j-j/wp/wp-je18/index/pdf.html （参照 2020-02-06）.

(b) 厚生労働省：厚生労働白書, https://www.mhlw.go.jp/stf/wp/hakusyo/kousei/18/ （参照 2020-02-20）.

(c) 経済産業省：通商白書, https://www.meti.go.jp/report/tsuhaku2019/whitepaper/2019.html（参照 2020-02-20）.

(d) 総務省：地方財政白書, https://www.soumu.go.jp/menu_seisaku/hakusyo/chihou/31data/2019data/mokuji.html（参照 2020-02-06）.

(e) 外務省：外交青書, https://www.mofa.go.jp/mofaj/fp/pp/page22/003299.html （参照 2020-02-15）.

(f) 人事院：公務員白書, https://www.jinji.go.jp/hakusho/pdf/index.html （参照 2020-02-22）.

(g) 警察庁：警察白書, https://www.npa.go.jp/hakusyo/r01/pdf/pdfindex.html （参照 2020-02-15）.

(h) 公正取引委員会：公正取引委員会年次報告, https://www.jftc.go.jp/soshiki/nenpou/h30.html （参照 2020-02-16）.

(i) 中小企業庁：中小企業白書, https://www.chusho.meti.go.jp/pamflet/hakusyo/2019/PDF/2019_pdf/mokujityuu.htm（参照 2020-02-20）.