文体変化と文体類似度を用いた文章の執筆者数推定

渡邉 充博† Eint Sandi Aung† 山名 早人‡,§

† 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保3-4-1
‡ 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保3-4-1
§ 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋2-1-2
E-mail: †‡{mwatanabe, esdaung, yamana}@yama.info.waseda.ac.jp

あらまし Web上では誰もが容易に情報を発信できる. こうした中, ブログやSNSの普及によって個人の情報発信の機会は一層増加し, 気軽かつ即座に文章を投稿できるようになった. 一方, 近年ではフェイクニュースと呼ばれる虚偽の情報を含んだニュースの拡散が問題となっており, 読者が内容の真偽を判断する必要性が高まっている. 本稿では,「信頼性の高い文章ほど編集に関わった人数が多い傾向にある」という点に着目し, 文章の執筆者数を信頼性の測定指標として利用する. 具体的には, 文体の類似度を用いて文章中の文体変化を検出し執筆者数を推定する手法を提案する. 評価実験では複数人による記述を含む文章に対して提案手法を適用し文体の変化点と執筆者数を推定した. 2人の異なる執筆者による記述を含む文章について, 執筆者数を推定し平均絶対誤差0.328人の結果を得た.

キーワード 情報信頼性, 文体分析

1. はじめに

Webの発展に伴って誰もが容易に情報を発信できるようになった. ブログやSNSの普及によって従来は情報を消費する立場にあった人々も情報の発信者となり, 多様な情報が即座に生成され入手できるようになった. 一方で掲載されている情報の正しさや有用性は必ずしも保証されていない. 近年ではフェイクニュースと呼ばれる虚偽の情報を含んだニュースの拡散が問題となっており, 事実であるニュースと比較してより早く広く拡散することが明らかになっている [1]. SNS上で拡散するという行為を通してユーザは誤情報の受信者に留まらず発信者にもなり得る. したがって情報の内容を見極めて真偽を判断する必要性は高まっている.

内容の真偽を判定するには, 関連する情報の知識が必要となり多くの手間と時間を要することから, 信頼性の自動推定が求められている. 信頼性推定自動化にはページ中の単語・文章の特徴やリンク関係が用いられる. 文章の執筆や編集の特徴に着目したものとしては, Wikipedia[1]の記事の編集の頻度 [2] [3]や, 削除されずに残る文章 [4]から品質の高い記事や文章の推定を行う研究がある. また, Wikipediaの記事の編集者数の分析を通して品質の高い文章ほど編集に関わった人数が多い傾向にあることが明らかにされている [5].

しかし, 編集の履歴を保持するWikipediaの記事と異なり一般の文章では何人によって執筆されたかは明らかでない. 文章の執筆者数を信頼性の測定指標として利用するためには, 執筆者数を推定することが必要となる. 執筆者数の推定を行う研究は少なく, 塩浦ら [6]の手法がわれわれの知る限りである. [6]では, 文章を分割し文体を基準にクラスタリングを行って人数を推定しているが, クラスタリングを用いているため文章内の文体の変化を定量的に扱うことができていない. そこで本稿では, 文体の類似度を用いて文章中の文体の変化を検出し執筆者数を推定する手法を提案する. 評価においては, 複数人による記述を含む文章に対して提案手法を適用し文体の変化点と執筆者数を推定する.

本稿は以下の構成を取る. 2節で情報の信頼性および執筆者数推定を扱う関連研究を述べ, 3節で文体の類似度から文体の変化を検出し執筆者数を推定する手法を提案する. 4節で提案手法の評価実験の結果および考察を述べる. 最後に5節でまとめを述べる.

2. 関連研究

2.1. 情報の信頼性と品質

情報の信頼性は人間が知覚する情報の品質である. Foggら [7]は情報の信頼性に関する従来の研究を統合して信頼性を次の4種類に整理した.

仮定的信頼性: 一般的な仮定に基づく信頼性
表面的信頼性: デザインや構造から知覚される信頼性
評判的信頼性: 第三者の評価に基づく信頼性
経験的信頼性: 過去の経験から判断される信頼性

Foggら [8]はWebサイト信頼性評価の際にユーザが着目する要素を調査することを目的に, 100件のサイトに対するユーザの合計2,440件のコメントを収集した. これらのうち46.1%のコメントではデザインに着目して評価が行われてお

[1] https://en.wikipedia.org/

り，ユーザは表面的な要素から信頼性を判断する傾向にあることが示されている．したがって，「情報の正しさや有用性を示す品質」と「知覚される信頼性の高さ」は必ずしも一致しないことがわかる．一方，「高品質の情報」は信頼性も高いことが示されている．

また，[9]ではWikipediaを対象として，専門家や一般のユーザによるWeb情報の品質評価が行われている．Wikipediaでは，ユーザによる推薦と投票から「秀逸な記事」と呼ばれる高品質な記事が選出されており，秀逸な記事と一般の記事の区分をもとに品質評価のための指標評価が行われている．品質評価においては，記事の差し戻し回数[2]や安定性[3]といった編集頻度のほか，質の高い記述は削除されにくいことから文章の残存率[4]が評価指標として提案されている．

2.2. 文章の編集と品質

Wilkinsonら[5]は，Wikipediaの各記事が扱うトピックの人気や注目されやすさが記事の品質に与える影響を調査するため，1,211件の秀逸な記事を含む150万件の記事の編集回数を分析した．まず，記事の人気度を示す指標として，閲覧回数と相関があるPageRankを用い，人気度と編集回数との関係を明らかにした．結果，PageRankの値が大きいほど(閲覧数の多い記事ほど)，編集回数と編集者数が多い傾向にあることが示されている．なお，編集回数と編集者数は記事の存在期間で正規化した値を用いている．また，同じPageRankの記事群内で比較した場合，秀逸な記事と一般の記事との間で編集回数および編集者数には有意な差が見られ，品質の高い記事ほど多くの編集が加えられていることが明らかになった．

一方で，品質評価を目的として執筆や編集に関わった人数を推定する取り組みはこれまでほとんど行われていない．塩浦ら[6]による手法がわれわれの知る限りである．[6]では，文章に対して品詞n-gramの頻度ベクトルを求め記述をクラスタリングすることで執筆者数を推定している．一定長のスライディングウィンドウにより文章を分割し，各ウィンドウ内の単語に対して品詞n-gramを取り，nを用いて重みを付けた頻度ベクトルを求める．続いてx-meansにより頻度ベクトルのクラスタリングを行っている．複数人による記述を含むウィンドウのベクトルはクラスタリング精度低下の要因となるため，クラスタリング精度の評価指標であるSilhouette係数に基づき一部のベクトルを削除する．残ったベクトルに対して再びクラスタリングを適用し，得られたクラスタの数を執筆者数としている．また，頻度ベクトルの重み付けに$\log(3n)$を用いると最も高いクラスタリング精度を示すことが報告されている．

2.3. 著者推定と文体分析

[6]で用いられた品詞のn-gramは，著者推定のタスクで用いられる手法である．著者推定では文章の特徴から文体を分析し著者の判別を行う．文章の特徴量は文を構成する単位別では文字，単語，文節，文，段落単位の特徴量に分けられる．各単位において用いられる特徴量は次のように分類される[10]．

文字:　文字種，n-gram，位置
単語:　品詞，語種，種類，単語長，接続位置，頻度
文節:　係り受け
文:　　種類，長さ
段落:　長さ，関連度

要約統計量としては，

- 全単語数に対する語彙数の比率
- 全単語数に対する特定頻度の語彙数の比率
- 特定の語彙の連続出現確率

が用いられ，統計量同士の比較には語彙の頻度ベクトルや語彙集合の類似度が用いられる．

[6]の手法はクラスタリングを用いているため，文章内におけるウィンドウの連続性が反映されない．また，文章内の文体の変化を定量的に扱うことができない．

3. 文体類似度を用いた文章の執筆者数推定手法の提案

本稿では，[6]における問題点（文書内におけるウィンドウの連続性が反映できない点，文体変化を定量的に扱えない点）を解決するため，品詞n-gramの出現頻度に基づく類似度を用いる手法を提案する．具体的には，文章をスライディングウィンドウで分割し，各ウィンドウ内に含まれる単語を品詞n-gramの出現頻度をもとに特徴量を付与する．そして，この特徴量が変化する位置を文体変化が発生した位置として検出し，執筆者数を推定する．提案手法は次の6つの処理から構成される．

1. 文章中の単語を品詞名に置き換え品詞列を取得する．
2. 品詞列をスライディングウィンドウにより部分品詞列に分解する．
3. ウィンドウ内の品詞列を半分に分割した前半部と後半部の文体の類似度を算出する．
4. 各ウィンドウについて3を実行し，類似度の低下点を抽出し，文体変化点の候補とする．
5. 各文体変化候補点においてより長いウィンドウを用いて類似度を再度算出し，候補を絞り込む．

6. 残った文体変化点の数から執筆者数を決定する.

3.1. 品詞n-gramの取得

執筆者数の推定を行う文章に形態素解析を適用し単語の品詞情報を取得する. 形態素解析器にはMeCab, 辞書にはNEologdを用い, MeCabが出力する品詞情報の2階層目までを取得する. 例えば, 名詞「雨」の場合は1階層目「名詞」, 2階層目「一般」となる. 句読点と助詞, 助動詞については, 品詞名に置き換えずに原文の表記を用いる. 具体的には, 「雨が降った. 」という文であれば, 含まれる単語のうち, 品詞情報が表3.1に示すものである単語は原文の表記を用いて, [名詞 一般], [が], [動詞 自立], [た], [.]という品詞列が得られる.

表3.1 原文の表記のまま用いる単語の品詞情報

1階層目	2階層目
記号	句点
記号	読点
助詞	(全て)
助動詞	(全て)

得られた品詞列をスライディングウィンドウによって切り出して部分品詞列を取得する. ここでは, 文章の一部分をウィンドウにより切り出すことで,「1つのウィンドウ内には最大2人による記述のみを含む」と仮定している. このとき, 各ウィンドウの記述について1人によるものか2人によるものかどうかを判別すればよいことになる. 一方, 2人を超える執筆者数は, ウィンドウをスライドさせることにより発見することが可能となる.

まず, 図3.1のようにウィンドウ内の品詞列を半分に分割する. 前半部と後半部でそれぞれ品詞列のn-gramを取って, 語彙の出現頻度のベクトルを作成する. nには1から3を用いる.

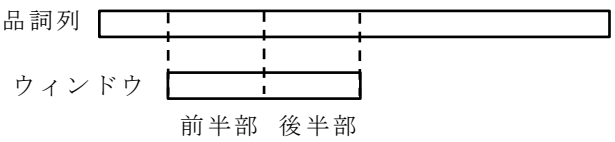

図3.1 ウィンドウの構成

前半部/後半部における出現回数を文章における出現回数で割ることにより正規化する. また, nが大きくなるにつれて当該語彙の発生確率は減ることから, nに応じた値で重みを付ける. [6]の結果から$\log(3n)$を乗算する. 全体の品詞列をw, 前半部または後半部の部分品詞列をs, $freq(v)$を語彙vの出現回数としたとき, 語彙vの出現頻度は式3.1で求められる.

$$freq = \log(3n)\frac{freq_s(v)}{freq_w(v)} \tag{3.1}$$

最後に, 前半部および後半部の各々の特徴ベクトルを, 各語彙に対応する出現頻度(式3.1)を用いたベクトル(品詞n-gramの頻度ベクトル)で表現する.

3.2. 文体変化点の推定

文体変化点を推定するため, ウィンドウ内の前半部と後半部の品詞n-gramの頻度ベクトル同士のコサイン類似度を求める. コサイン類似度はベクトル$\boldsymbol{a}, \boldsymbol{b}$を用いて式3.2で表される.

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} \tag{3.2}$$

ウィンドウごとに当該ウィンドウ内の前半部($\boldsymbol{a}$)と後半部($\boldsymbol{b}$)の間で類似度を求めることにより, 文章中の類似度の変化を得る. このとき, $\boldsymbol{a}$と$\boldsymbol{b}$で文体が異なるほど類似度は小さくなるため, あるウィンドウwin_iにおける類似度sim_iがウィンドウwin_{i-1}より小さく, かつ, ウィンドウwin_{i+1}よりも小さい時, ウィンドウwin_iの中心を文体変化発生の候補点とする. ただし, 1つのウィンドウ内には最大2人による記述を含むと仮定しているため, ウィンドウの半分の長さを候補点同士の最小間隔とし, 間隔が最小間隔を下回る点は類似度が低い方の候補点を残す.

次に, 得られた複数の候補点から絞り込みを行う. 図3.2のように候補点C_iを中心として前後の候補点C_{i-1}とC_{i+1}のうちC_iにより近い点を端点とするようにウィンドウ長を伸ばす. ウィンドウ長を伸ばすことで類似度算出時に利用できる単語数が増える. 伸ばしたウィンドウwin_{ext}の長さは式3.3で求められる.

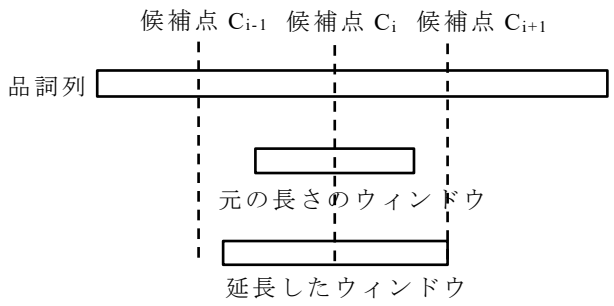

図3.2 ウィンドウの延長

$$len(win_{ext}) = 2 \cdot \min\left(|C_i - C_{i-1}|, |C_i - C_{i+1}|\right) \tag{3.3}$$

品詞列に対してwin_{ext}をスライドさせ3.1項と同様に部分品詞列を取得し類似度を求める．あるウィンドウ$win_{ext,j}$の類似度がウィンドウ$win_{ext,j-1}$の類似度より小さく，かつウィンドウ$win_{ext,j+1}$の類似度より小さい時，ウィンドウ$win_{ext,j}$の中心を文体変化発生の候補点$C_{ext,j}$とする．候補点C_iの位置から一定の範囲内に$C_{ext,j}$が存在しない場合，C_iを削除する．

さらに，C_iから一定の範囲内に$C_{ext,j}$が存在する場合でもwin_{ext}に基づく類似度の平均値$\overline{sim_{ext}} = 1/n \cdot \sum_n sim_j$（$win_{ext}$がn個ある場合）と，候補点の類似度$sim_{ext,j}$の差$\overline{sim_{ext}} - sim_{ext,j}$があらかじめ設定した閾値より小さい場合$C_i$を削除する．残った点を文体の変化点とし，執筆者が変わったと判断する．文体変化点の数が推定される執筆者数となる．

4. 実験と評価

本実験では予備的評価として，2 人によって記述された文章であるかどうかを判定できるか検証した．4.1 項で用意した文章に対して 3 節の手法を適用し，その精度を検証する．

4.1. データセット

データセットを作成するにあたり，一人のみによって記述された文章を用意する．小説の文章は一人のみによって書かれているとみなして，青空文庫[2]に収録されている作品を用いた．異なる執筆者の小説作品の文章を切り出して連結することで複数人によって書かれた文章を生成する．文章の切り出しは文単位で実行される．各作品の文章の先頭から500単語以上を含むように文群を抽出した．4.2項の実験に対しては10人の執筆者の各5作品，合計50作品，4.3項の実験に対しては10人の執筆者の各10作品，合計100作品を用いた．文群から2個ずつ選択して連結し一つの文章とすることで2人によって記述された文章を生成した．

4.2. 閾値の決定

同一文章内での各ウィンドウにおける類似度の平均値との差をもとに閾値を決定する．閾値の決定には，10人の執筆者による各5作品，合計50作品を用いた．各文章の先頭500単語以上を含むように文を切り出すと，平均で520.28単語，19.34文となり，同一執筆者の文章が100組，異なる執筆者の組み合わせが1,125組となった．異なる執筆者の組み合わせ1,125組から無作為に選択した100組を用いた．各文章においてスライディングウィンドウで求めた類似度から候補点を抽出し，文章内の各ウィンドウの平均類似度と候補点C_iにおける類似度との差$\overline{sim} - sim_i$の値が，閾値以上の場合に文体変化点である，閾値未満の場合に文体変化点でないと推定する．候補点のうち文体変化点と推定した点から60単語以内に実際の文体変化点が存在する点の数をTP，そうでない点の数をFPとし，候補点のうち文体変化点でないと推定した点から60単語以内に実際の文体変化点が存在する点の数をFN，そうでない点の数をTNとして計数を行い，式4.1より適合率pを文章ごとに求める．

$$p = \frac{TP}{TP + FP} \tag{4.1}$$

d(=100)個の文章に対して p の平均を求め，$\bar{p} = p/d$が最大となる値を閾値と決定した．ウィンドウのスライド幅を 30 単語とし，表 4.1 で設定したウィンドウ長ごとに閾値を決定した．表 4.1 の結果を線形補間して任意のウィンドウ長における閾値を得た．

表 4.1 ウィンドウ長ごとに決定した閾値

ウィンドウ長[単語]	閾値
100	0.009
200	0.013
300	0.013
400	0.008
500	0.007

4.3. 文体変化点の推定

続いて1人もしくは2人によって記述された文章に対して文体変化点および執筆者数を推定した．4.2項で用いた執筆者とは異なる10人の各10作品，合計100作品を用いた．各文章の先頭500単語以上を含むように文を切り出すと，平均で517.66単語，18.69文となり，同一執筆者の文章が450件，異なる執筆者の組み合わせた文章が4,500件となった．

執筆者を区別するために必要な文字数は，これまでのわれわれの知見 [11]に基づき少なくとも100文字以上が必要であることが分かっている．そこで本実験ではウィンドウ長を200，300，400単語の3種類に設定した．ウィンドウのスライド幅は30単語とした．

同一の執筆者の組み合わせの文章450件と，異なる執筆者の組み合わせの文章4,500件から無作為に選択した450件について，推定執筆者数と実際の執筆者数が一致した確率(条件はTP+FP=TP+FN)および，推定執筆者数と実際の執筆者数との平均絶対誤差MAEを求めた．また，異なる執筆者の組み合わせの文章に対しては推定執筆者数と実際の執筆者数が一致し，かつ文体変化の推定点が実際の変化点の前後60単語以内にあった確率(条件はFP=FN=0)を求め

[2] https://www.aozora.gr.jp/

た. 結果を表4.2に示した.

表4.2 執筆者数の推定結果

	同じ執筆者		異なる執筆者		
ウィンドウ長[単語]	執筆者数を正しく推定できた確率	MAE[人]	執筆者数を正しく推定できた確率	左記に加え，文体変化点を正しく認識できた確率	MAE[人]
200	0.556	0.684	0.271	0.091	0.797
300	0.508	0.573	0.478	0.313	0.533
400	0.424	0.578	0.671	0.467	0.328

ウィンドウ長が400単語の場合，1人の執筆者によって記述された文章のうち42.4%が執筆者1人，57.6%が2人以上と推定され，推定人数のMAEは0.578人であった. また，2人の執筆者による文章のうち67.1%が執筆者2人と推定され，32.9%が執筆者1人もしくは3人以上と推定された. ただし執筆者数が正解し，かつ文体変化点を正しく推定できていたのは46.7%であり，MAEは0.328人であった.

ウィンドウ長が長いほど文体変化点と執筆者数の推定精度の向上が見られた. ウィンドウが短いほど，実際の文体変化点に対応しない点が候補点として抽出されやすくなるため，ウィンドウ長が長いときより候補点の絞り込みで削除される点が増える. つまり文体変化点が0個と推定される場合が多くなるため執筆者数1人の文章に対しては短いウィンドウほど執筆者数が正しく推定される確率は高くなった.

5. まとめ

本稿では文体の類似度を用いて文章中の文体の変化を検出し執筆者数を推定する手法を提案した. 文章に対して品詞n-gramの頻度をもとに特徴量を定義し，文章の類似度を算出し文体の変化点および執筆者数を推定した. 2人の異なる執筆者による記述を含む文章に対して，ウィンドウ長を400単語とした場合，文体変化点および執筆者数を正しく推定できた確率として46.7%，推定執筆者数の平均絶対誤差0.328人という結果が得られた. 今後は推定精度向上に取り組み3人以上の執筆者の記述を含む文章に対応させるほか，Wikipediaの記事のような執筆者数が既知である実際の文章に対して適用，検証することや評価指標として信頼性推定手法に組み込んで利用することを考えている.

謝 辞

本研究の一部は，科学研究費助成事業17KT0085によるものである.

参 考 文 献

[1] S. Vosoughi, D. Roy and S. Aral, "The spread of true and false news online," *Science,* vol.359, issue 6380, pp.1146-1151, 2018.

[2] F. B. Viegas and M. Wattenberg, "Talk Before You Type: Coordination in Wikipedia," in *Proc. of HICSS '07*, pp.78-87, 2007.

[3] P. Dondio, S. Barrett, S. Weber and J. M. Seigneur, "Extracting Trust from Domain Analysis: A Case Study on the Wikipedia Project," in *Proc. of ATC '06*, vol.4158, pp.362-373, 2006.

[4] B. T. Adler, L. d. Alfaro, I. Pye and V. Raman, "Measuring Author Contributions to the Wikipedia," in *Proc. of WikiSym '08*, no. 15, pp. 1-10, 2008.

[5] D. Wilkinson and B. Huberman, "Cooperation and Quality in Wikipedia," in *Proc. of WikiSym '07*, pp.157-164, 2007.

[6] 塩浦尚久，山名早人，"日本語の文章を対象にした執筆者人数推定", DEIM Forum 2019 論文集, B5-1, 2019.

[7] B. J. Fogg and H. Tseng, "The Elements of Computer Credibility," in *Proc. of CHI '99,* pp.80-87, 1999.

[8] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford and E. R. Tauber, "How do users evaluate the credibility of Web sites? a study with over 2,500 participants," in *Proc. of DUX '03*, pp.1-15, 2003.

[9] T. Chesney, "An empirical examination of Wikipedia's credibility," *First Monday,* vol.11, no.11, 2006.

[10] 浅石卓真，"テキストの特徴を計量する指標の概観"，日本図書館学会誌，vol.63, no.3, pp.159-169, 2017.

[11] S. Okuno, H. Asai and H. Yamana, "A Challenge of Authorship Identification for Ten-thousand-scale Microblog Users," in *Proc. of IEEE Int'l Conf. on BigData*, pp.52-54, 2014.