大規模候補者群に対する著者推定手法の提案と評価

井上 雅翔† 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: †m.inoue@yama.info.waseda.ac.jp, ‡yamana@yama.info.waseda.ac.jp

あらまし 近年，インターネットに投稿された文章を対象とした，大規模人数の候補者を対象とした著者推定が求められている．そこで本稿では，1 万人レベルの大規模候補者群を対象としても利用可能な著者推定手法を提案する．具体的には，著者推定のために，著者らが提案している品詞列の頻度を用いる手法と既存手法である文字列の頻度を用いる手法とを併用する著者推定手法を提案する．提案手法は，推定対象文章の話題変化に頑健であり，既存手法より高精度な著者推定，もしくは既存手法よりも高速な推定を実現する．さらに，本稿では大規模候補者群に対する著者推定に適した評価手法を併せて提案し，提案手法と既存手法の比較を行った．10,000 人の候補者から著者を推定する評価実験を行った結果，提案手法は既存手法と同等精度を保ちながら既存手法よりも 4.9 倍高速な著者推定を実現した．また，著者推定に必要な計算量増加を許し，複数手法を併用する手法では，7,260 人の候補者から著者を推定する評価実験を行った結果，既存手法よりも 15%の著者推定精度の向上に成功した．

キーワード 著者推定，テキストマイニング，大規模データ

1. はじめに

既存の著者推定手法[1][2][3]は小説などの文学的文章における著者推定を実現し，近年ではインターネットに投稿された日本語の文章を対象とした著者推定に応用[4][5][6][6]されている．既存の著者推定手法を用いて，インターネットに投稿された文章の著者を推定する際には，文学的文章の著者を推定する場合とは異なり，大規模人数の著者候補者群に対して著者を推定する必要がある．なぜならば，文学的文章の著者は限定されているが，インターネットに文章を投稿する著者は不特定多数であり，少人数に限定できないためである．

既存の著者推定手法を用いて大規模候補者群に対する著者推定を行うと，文学的文章を対象とした著者推定の結果に比べて推定精度，及び推定処理速度が低下する．これは，大規模候補者群に対する著者推定において以下の 3 つの問題に起因している．

1. 文体類似著者の高頻出
2. 同一話題文章収集困難化
3. 推定処理時の計算量増加

1 つ目は，候補者群の中で類似した文体を持つ者が多く発生する問題である．文体とは，文章を書く際に現れる，個人毎に異なる癖である．著者推定は推定対象文章の著者と文体が同じ候補者を探すことで行われる．しかし，候補者群の中で文体が類似する者が多く現れると，実際の著者と文体が類似する当該著者以外の候補者を誤って実際の著者と判定してしまう．この時，著者推定は失敗するため推定精度が低下してしまう．2 つ目は，著者推定で用いる候補者毎の文章を，各々同一話題にして収集できなくなる問題である．これは，1 つの話題について大量に文章を書く著者は一般的に少ないために起こる．著者推定で用いる候補者の文章が同一話題でなく相違話題となる時，著者推定精度は低下する．これは，我々の先行研究である文献[8]で示されている．3 つ目は，推定対象文章の著者を推定するときに処理しなくてはならない計算量が増加する問題である．これは，推定対象文章ごとにすべての候補者に対して著者推定の処理をしなくてはならないためである．

本稿では，大規模候補者群に対する著者推定で発生する推定精度，及び推定処理速度低下に対応すべく，上記の 3 つの問題に対応した著者推定手法を 2 つ提案する．

最初の手法は 3 つの問題すべてに対応した手法である．大規模候補者群に対して相違話題文章の著者を高精度かつ高速に推定する．2 つ目の手法は推定処理時の計算量増加を許すことで，1 つ目の手法よりも高精度に著者を推定する．評価実験では，我々が新たに提案する大規模候補者群に対する著者推定評価手法を用い推定精度を評価すると共に，推定処理速度を評価する．なお，同一話題文章収集困難化の問題を再現するため，本評価実験で扱う推定対象文章は相違話題であるものも用いる．当評価実験のため，ニフティサーブにおけるフォーラムの電子掲示板に投稿された文章[9]を用いる．この電子掲示板は所属するフォーラムが異なると話題も異なるため，フォーラムの相違によって相違話題文章を収集することができる．

本稿は以下の構成をとる．まず 2 節では，著者推定研究で取り扱われてきた著者推定タスクについて述べる．次の 3 節では，既存の著者推定手法について述べる．続く 4 節では，本稿で提案する著者推定手法について述べる．そして，5 節にて既存手法と提案手法とに対する評価実験の方法と結果について述べる．最後の 6 節で本稿をまとめる．

2. 著者推定タスク

著者推定とは，推定対象文章における文体の特徴から，その文章の著者を推定することである．推定対象文章とは，著者を推定する対象となる，著者不明の文章のことである．なお，本稿では日本語の推定対象文章を対象とした著者推定を取り扱う．著者推定で取り扱う文体の具体例としては，語彙の選び方，文章の構成方法，句点，読点の打ち方が挙げられる．著者推定は文学研究[10][11][12]で行われてきたが，近年ではテ

キストマイニング技術を用いた著者推定の手法[1][3][6][6]が提案されている．これらの手法は計算機上で容易に実装可能であることから，インターネットに投稿された文章の著者推定に応用[4][5]されている．テキストマイニングによる著者推定手法の研究では，著者推定手法をもって著者推定タスクを行い，この結果によって当手法の評価を行う．本節では，著者推定タスクについて，その内容と結果からの評価方法について述べていく．

2.1. 著者推定タスクの分類

Stamatatos[13]は著者推定タスクを Profiled-Based Approach(PBA)と Instance-Based Approach(IBA)の 2 種類に分類した．本稿では，大規模候補者群に対する著者推定を行うため，PBA による著者推定タスクを取り扱う.これは,大規模候補者群に対する著者推定では，IBA による著者推定タスクに 2.1.2 で示す問題があるためである．

2.1.1. PBA 及び IBA による著者推定タスク

PBA による著者推定タスクでは，事前に用意されている候補者の文章群と,推定対象文章を順に比較する．比較された候補者群の中から，最も推定対象文章の著者と文体が類似する候補者を得ることで，各著者推定手法は著者推定を行う．PBA に分類される著者推定タスクは，松浦ら[1]，安形ら[2]，中島ら[6]，及び我々の先行研究[8]が取り扱っている．

一方で，IBA による著者推定タスクでは，機械学習により各候補者の文章群を学習し，推定対象文章を各候補者のいずれかに分類する．推定対象文章の分類先となる候補者を得ることで，各著者推定手法は著者推定を行う．IBA に分類される著者推定タスクは，金ら[3]，坪井ら[6]が取り扱っている．

2.1.2. IBA による著者推定タスクの問題点

大規模候補者群に対する著者推定における IBA の著者推定タスクでは，当該著者推定タスクにおける機械学習が上手く機能しない．これは，IBA の著者推定タスクで用いる学習データが不均衡データとなるために起こる[14].

不均衡データとは，正例と負例の数に極端な差がある学習データを示す．IBA における著者推定タスクでは，学習データ中の文章群を，特定の 1 人の候補者の文章である正例，及び，それ以外の複数候補者の文章である負例の 2 つに分割する．しかし，一般に負例を集めることは容易であるが，正例を多く集めることは困難である．このため，IBA における著者推定タスクでは，正例と負例の数に差が生まれ，学習データは不均衡データとなる．

不均衡データに対処するため，正例の数に合わせて負例の数を減らしたり，負例の数に合わせて正例の数を多くしたりする対策が考えられる．しかし，前者の方法では学習が十分にできない問題が生じる.一方で，後者の方法を講じることも難しい．これは，候補者ごとに集められる文章は数万文字の大量文章でなくてはならないが，このような文章を 1 人の候補者に対し多く集めることは困難であるからである．

2.2. PBA による著者推定タスクの流れ

手順 1）学習データとテストデータの収集

学習データとは，著者が既知である文章群のことを指す．テストデータとは複数の推定対象文章を指す．ただし，著者推定タスクでは，推定したテストデータ中の文章の著者と実際の著者が同じであることを確かめるため，テストデータ中の文章の著者も事前にわかっているものを用いる．また，テストデータ中の文章の著者は，学習データにおけるいずれかの文章の著者と同一であるとする．このような条件の下，著者推定の候補者群となる著者を決定した後，候補者ごとに学習データとテストデータの 2 種類の文章を集める．

手順 2) 各文章の文体定量化

手順 1 で収集された学習データ及びテストデータ中のすべての文章に対して文体定量化を行う．文章の文体定量化とは，その文章の著者が持つ文体を，当該文章を用いて数値ベクトルに定量化することである．文章に対する文体の定量化方法は，各著者推定手法によって異なる．

手順 3）各文章間の文体相違度計算

テストデータ中の文章ごとに，学習データ中の各文章との間の文体相違度をすべて計算する．2 つの文章間の文体相違度とは，各文章の著者の文体がどれほど異なるかを定量化したものである．2 つの文章間の文体相違度は，手順 2 で得られる定量化された文体を用いて算出される．文体相違度をどのように算出するかは，各著者推定手法によって異なる．

手順 4）文体類似度順位の算出

テストデータ中の文章ごとに文体類似度順位を算出する．文体類似度順位とは，文体相違度の低い順に候補者群を並び替えたとき，推定対象文章の著者が何位に順位付けされたかを表す．テストデータ中の各文章を推定対象文章とするとき，学習データ中の各文章の著者との文体類似度は手順 3 で求められている．

手順 5）著者推定手法の評価

手順 4 で得られたテストデータ中の各文章に対する文体類似度順位に基づいて，手順 2 及び手順 3 で用いた著者推定手法の評価を行う．得られた文体類似度順位からどのように著者推定手法を評価するかは，著者推定手法評価方法によって異なる．

2.3. 評価方法

本稿では，著者推定タスクから得られる結果の評価に，既存の著者推定研究で行われる評価方法とは異なる，新たな評価方法を提案する．

既存の著者推定研究[1][2][3][6][6][8]で行われる著者推定手法の評価は，2.2 で述べた著者推定タスクの手順 5 において，テストデータ中の文章群の中で文体類似度順位が 1 位となる文章の割合が用いられてきた.これは，テストデータ中の各文章に対して著者推定を行うとき，著者推定タスクの手順 4 で並び替えられる候補者群において 1 位となる候補者を推定対象文章の著者であると推定するためである．

本稿では，情報検索で用いられている平均逆順位と，Mean Top-k Recall による評価方法を用いる.これらは，著者推定タスクによって候補者群を並び替えた時，上位 k 件に実際の著者が含まれるかどうかを評価するために用いる．当該の評価に対して，平均逆順位を指標として用いることができるが，これだけでは「上位 k 件の中に正解が含まれる確率」を的確に表現できないため，Mean Top-k Recall を用いる．具体的には，Mean Top-k Recall よって，上位 k 件の中に実際の著者が含まれる確率を，k をパラメータとしてグラフ化する．

著者推定タスクの結果に対する評価に，本稿で提案する評価方法を用いるのは，著者推定タスクにおける候補者群の並び替えにおいて，実際の著者が一位に順位付けされているかだけでなく，上位に順位付けされているかを評価するためである．これは，誤った推定をしない著者推定手法が存在しない以上，推定結果を実用するためには人手による確認が要求される.特に，

推定精度低下が顕著となる大規模候補者群に対する著者推定では，人手による確認の要求が顕著となる．推定結果を人手で確認するためには，並び替えられた候補者群の 1 位 1 人だけではなく，上位複数人を相互に比較しながらの確認が行われる．よって，本稿で行う大規模候補者群に対する著者推定に対しては，複数人の推定結果を得ることを考慮した，本稿で提案する評価方法が用いられる．

3. 従来の著者推定手法

著者推定タスクにおける文体定量化や文体相違度算出に対して，多数の著者推定手法が提案されている．これらの著者推定手法は，文字列の頻度を用いる手法か，品詞列の頻度を用いる手法かで 2 分することができるが，各々で固有の問題を持っている．

3.1. 既存手法

3.1.1. 松浦らの手法

松浦ら[1]は，著者推定タスクにおける文体定量化に文字 n-gram 頻度分布を用いることで，日本語の文学的文章を対象とした著者推定手法を提案している．文字 n-gram とは，文章中に現れる連続する n 文字の文字列のことである．また松浦らは，著者推定タスクにおける文体相違度計算に，Tankard の文章間相違度計算方法[15]を用いている．

松浦らが提案する文字 n-gram 頻度分布による文体定量化は，文章pにおける文字 n-gram xの出現確率分布関数$P_p(x)$を得ることで行う．$P_p(x)$は，文字 n-gram xが文章pで出現する確率を表すものであり，文章pに現れるすべての文字 n-gram に対する文章pで出現する文字 n-gram xの数の割合である．

松浦らが用いた Tankard の文章の相違度計算は，文章pと文章qとの文体相違度 $Dissim$ を，$P_p(x)$を用いて以下のように計算することで行う

$$Dissim(p,q) = \sum_{x \in C} \left| P_p(x) - P_q(x) \right| \quad (1)$$

なお，Cは文章p,q中に現れるすべての文字 n-gram の集合である．ただし，C中の要素は，重複して存在することが許されない．また，文体相違度 $Dissim$ は，その値が小さいほど 2 つの文章p,qの文体が似ていることを表している．

3.1.2. 安形らの手法

安形ら[2]は，著者推定タスクの文体相違度計算に，ZIP 圧縮プログラムを用いることで，日本語の文学的文章を対象とした著者推定手法を提案している．当該手法は文章に対する文体定量化の処理を行う必要がない．当該手法の文体相違度計算方法は以下の式によって行う．

$$Sim(p,q) = 2 \cdot \frac{LZ_p + LZ_q}{L_{p+q}} - \frac{LZ_{p+q} + LZ_{q+p}}{L_{p+q}} \quad (2)$$

LZ_pとは， ZIP 圧縮プログラムによって文章pを圧縮したバイト列長である．また，LZ_{p+q}とは，文章pの後ろに文章qを連結した文章に対するLZ_pの値である．L_{p+q}とは，文章pとqを連結させた文章の文字列長を表す．式(5)の文体類似度 Sim は，その値が大きいほど 2 つの文章p,qの文体が似ていることを表している．

3.1.3. 中島らの手法

中島ら[6]は，著者推定タスクにおける文体定量化に品詞 n-gram 頻度分布を用いることで，日本語のブログ記事を対象とした著者推定手法を提案している．品詞 n-gram とは，文章を品詞タグ列に変換したとき，その品詞タグ列の中に存在する n 個の連続した品詞タグ順列を指す．品詞タグとは，「名詞」や「動詞」などの形態素における品詞名を示す．また中島らは，著者推定タスクにおける文体相違度計算手法として，ピアソンの積率相関係数を用いた手法を提案している．

中島らが提案する文章中の文体定量化は，文章p中における品詞 n-gram xの生起回数d_{px}の集合D_pを得ることで行う．なお，品詞 n-gram を得るために，文章pを形態素解析によって形態素列に変換し，各形態素をその品詞タグに変換することで品詞タグ列を得る．

中島らが提案する著者推定タスクにおける文体相違度計算では，文章p,qについてのD_pだけでなく，C_{pq}, a_pを用いる．C_{pq}は，文章pと文章qの各々に存在するすべての品詞 n-gram の和集合である．a_pは，文章pを構成する記事の数である．記事とは，電子掲示板における 1 つの記事や，1 件の電子メールのように，一度に投稿する文のまとまりを指す．中島らは，C_{pq}, D_p, a_pを用いることで 2 つの文章p,qにおける文体相違度 $Dissim$ を以下のように定義している．

$$Dissim(p,q) = \frac{\sqrt{\sum_{i \in C_{pq}} (f_{pi} - \overline{f_{pq}})^2} \sqrt{\sum_{i \in C_{pq}} (f_{qi} - \overline{f_{qp}})^2}}{\sum_{i \in C_{pq}} (f_{pi} - \overline{f_{pq}})(f_{qi} - \overline{f_{qp}})} \quad (3)$$

$$\overline{f_{pq}} = \frac{\sum_{i \in C_{pq}} f_{pi}}{|C_{pq}|} \quad (4)$$

$$f_{pi} = \begin{cases} 0.4 \ (f'_{pi} > 0.4) \\ f'_{pi} \ (f'_{pi} \leq 0.4) \end{cases} \quad (5)$$

$$f'_{pi} = \frac{d_{pi}}{a_p} \quad (6)$$

文体相違度 $Dissim$ は，その値が小さいほど 2 つの文章p,qの文体が似ていることを表す．

3.2. 既存手法の問題点

3.2.1. 既存手法の分類

既存の著者推定手法は，文字列の頻度を用いる手法か，品詞列の頻度を用いる手法かの 2 つに分類できる．

文字列の頻度を用いる著者推定手法とは，著者推定で比較する文体を，推定で用いる文章中における各文字列の頻度によって定量化する方法である．当該手法には，松浦らの手法や安形らの手法が該当する．松浦らの手法は，文体定量化において文字 n-gram 頻度分布を用いることで，すべての連続する n 文字の頻度を要素とした数値ベクトルで文体定量化を行なっている．安形らの手法は，文体相違度計算に Zip 圧縮を用いることで，文章中の頻出する文字列に基づいて文体相違度を算出する．安形らの手法は，文体相違度を算出する 2 文章において共通して頻出する文字列が多いほど，結合された当該 2 文章に対する Zip 圧縮の圧縮率が大きくなる特性を利用している．

品詞列の頻度を用いる著者推定手法とは，著者推定で用いる文章中における各品詞列の頻度によって，文体を示す方法である．当該手法には，中島らの手法が該当する．中島らの手法は，連続する n 個の品詞を 1 品詞列として，各品詞列の頻度を要素とした数値ベクトルで文体定量化を行なっている．この各品詞列の頻度は，当該手法が文体定量化において構成する品詞 n-gram 頻度分布の各要素によって得られる．

3.2.2. 既存手法の各分類に対する問題点

文字列の頻度を用いる手法，及び品詞列の頻度を用いる手法によって行われる大規模候補者群に対する著者推定では，「推定処理時の計算量増加」または「類似文体著者の高頻出」の問題が顕著化してしまう．

文字列の頻度を用いる手法によって行われる著者推定タスクでは，推定処理時の計算量増加が顕著化する．これは，著者推定タスクの文体相違度計算ですべての文字列を比較することによって，2 文章間の文体を比較するためである．文字列の多様性は品詞列の多様性より大きいため，そのすべてについて比較処理を行う当該手法は，その計算量が比較的大きくなってしまう．よって，当該手法による著者推定では，推定処理の計算量増加が顕著化する．

品詞列の頻度を用いる手法によって行われる著者推定タスクでは類似文体著者の高頻出による問題が顕著化する．これは，文字列よりも種類の少ない品詞列では,個々に異なる文体を表現しきれないためである．表現しきれない文体の違いの分，得られる 2 つの文体の違いは小さくなる．このとき，文体が類似する 2 人の著者を判別することがより困難となり，著者推定精度は低下してしまう．

4. 提案手法

4.1. 品詞タグ,文字混合要素列の頻度を用いる手法

3.2 で述べた問題に対応するため，中島らの手法を改良した著者推定手法を提案する．具体的には，中島らの手法で文章中におけるすべての品詞 n-gram についてD_pを算出したのに対して，提案手法は文章中におけるすべての品詞タグ・文字混合 n-gram についてD_pを算出する．品詞タグ・文字混合 n-gram とは，文章を文字または品詞タグの羅列に変換した時に，当該羅列中に存在する n 個の連続した要素順列を指す．

文章を文字または品詞タグの羅列に変換するために以下の手順をとる．まず，形態素解析器を用いて文章を形態素に分割する．なお，形態素解析器は Sen[16]を用いている．次に,「動詞」「接続詞」「記号」「副詞」「形容詞」「感動詞」の形態素については，文字列をそのまま採用し，これら 6 種類の品詞以外については，品詞タグを用いる．以上の手順によって文章から品詞タグ及び文字の羅列に変換する例を，図 1 に示す．

当該手法を用いることで，3.2 で述べた問題が顕著化しにくく，さらに同一話題文章収集困難化の問題も顕著化しにくい著者推定を実現できる．3.2 で述べた問題に対応できるのは，文字・品詞タグ混合要素列の種類が，文字列より少なく，品詞列より多いためである．これは，当該手法における文体定量化において文章中の一部の文字列を品詞列に変換したために，当該要素列中の要素の種類が，文字より少なく品詞より多いからである．同一話題収集困難化の問題が顕著化しないのは,「動詞」「接続詞」「記号」「副詞」「形容詞」「感動詞」の 6 つの形態素の各文字列の頻度は，同一著者の 2 文章で相違話題となっても変化しにくいためである．よって，同一著者の 2 文章の間で相違話題となっても，文字・品詞タグ混合要素列の各頻度は変化せず，この方法で定量化された文体は話題変化に頑健となる．つまり当該文体を比較して行う提案手法による著者推定は，同一話題文章収集困難化に伴う著者推定で用いる文章の話題が変化しても，高精度に著者推定が可能となる．

4.2. 多手法併用手法

当手法は，n個の著者推定手法を併用することで，文体類似著者の高頻出，及び同一話題文章収集困難化に伴う推定精度低下に対応する．具体的には，任意のn個の手法$m_i(i=1..n)$が著者推定タスクの手順 3 で算出する文体相違度$Dissim_{m_i}(i=1..n)$を併用し，新たな文体

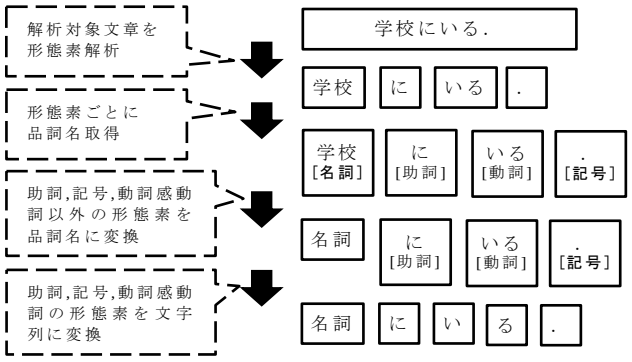

図 1　文章から文字・品詞タグの羅列に変換する例

相違度 *MergeDissim* を算出する．文体相違度$Dissim_{m_i}$は，本稿おける式(1),式(2),式(3)に相当する．*MergeDissim* は，既存の著者推定手法と同様に手順 4 の文体類似度順位の算出で用いる．文体相違度 *MergeDissim* の算出は，既存手法における文体相違度の算出と同様に文章p,qに対して行われる．ただし，当該文体相違度算出には，手順 1 で収集する学習データDを用いている．また，文体相違度計算の対象となる文章p,qについては，文章pを推定対象文章となるテストデータ中の文章とし，文章qを$q \in D$が成り立つ学習データ中の文章であるとする．以上を踏まえ，文章p,qに対する文体相違度 *MergeDissim* の算出は，以下の式で行う．

$$MergeDissim(p,q) = \sum_{i=1}^{n} a_i \frac{Dissim_i(p,q) - avr_i(p,D)}{vrp_i(p,D)} \quad (7)$$

$$avr_i(p,D) = \frac{1}{|D|}\sum_{d\in D} Dissim_i(p,d) \quad (8)$$

$$vrp_i(p,D) = \sqrt{\sum_{d\in D}\left(Dissim_i(p,k) - avr_i(p,D)\right)^2} \quad (9)$$

文体相違度 *MergeDissim* は，n個の手法で算出される文体相違度について各々偏差値を求め，これらの和によって算出される．文体相違度ごとに求められる偏差値とは，学習データD中の各文章と推定対象文章pとの間におけるすべての文体相違度に対する，文章p,q間の文体相違度の偏差値のことである．この偏差値を用いる文体相違度 *MergeDissim* によって，手法ごとに定数a_iに従う正しい重み付けを与えた文体相違度$Dissim_{m_i}(i=1..n)$の併合が可能となる．なおa_iは，当該手法で用いる各著者推定手法$m_i(i=1..n)$に対し事前に設定する定数である．

当手法に対してよりも，1 手法に対して存在する不得意推定対象文章の数のほうが多くなるため，1 つの手法だけを用いるよりも当手法は高精度に著者推定が可能となる．不得意推定対象文章とは，手法ごとに存在している特定の著者推定対象となる文章である．同一著者の 2 文章において文体相違度が小さくなる事実と相反して，この不得意推定対象文章を含む当該 2 文章の文体相違度は大きくなる．よって，ある著者推定手法に対する不得意推定対象文章が少ないとき，当該手法による著者推定は上手くいきやすく，その推定精度は高くなる．文体相違度 *MergeDissim* を用いる当手法は，その併用するn個の手法のいずれかに対して不得意推定対象文章とならない文章は，当該手法に対しても不得意推定対象文章とならない特徴を持つ．よって，1 手法に対してよりも，文体相違度 *MergeDissim* を用いた著者推定手法のほうが，不得意推定対象文章の数は少ない．

5. 評価実験

評価実験では，既存手法及び本稿で提案した手法を用いて，大規模候補者群に対する著者推定を行い，各手法の評価を行う．具体的には，2.2 で述べた著者推定タスクを行い，その結果に基づいて 2.3 で述べた，本稿で提案する手法評価を行う．また，推定処理時の計算量増加による推定処理速度低下の問題に対して，著者推定タスクの手順 3 における処理時間も併せて評価する．

5.1. 評価実験方法

5.1.1. データセット

本項では，本評価実験で使用したフォーラムの電子掲示板[9]について，その概要と特長を述べる．

フォーラムとは，ニフティサーブにおけるサービスの１つである．ニフティサーブとは，ニフティ株式会社が 1987 年から 2006 年まで運営していたパソコン通信サービスである[1]．ニフティサーブでは，フォーラムと呼ばれる場所を提供し，ネットワークを介して分野ごとにコミュニケーションを取ることができた．フォーラムの中には電子掲示板が存在しており，電子掲示板に文章を投稿し合うことで不特定多数の人と会話をすることができた．つまり，電子掲示板の所属するフォーラムが変わることで，そこに投稿される文章で話される分野が変わることになる．文章中で話される分野が異なると，その話題も異なるため，複数のフォーラムにおける電子掲示板に投稿される文章を用いて，異なる話題の文章群を用いた著者推定が可能となる．

現在，ネットワーク上に文章を投稿できるサービスは数多く存在するが，同様のサービスであるフォーラムにおける電子掲示板とは異なる点がある．それは，フォーラムの電子掲示板では，投稿文章に付与されるユーザ ID が，文章投稿者と 1 対 1 で高確率に対応するという特徴である．ユーザ ID とは，文章の投稿者を区別するために文章投稿者に割り振られる，個体識別符号のことである．ネットワーク上に文章を投稿できる他のサービスにもユーザ ID が存在するが，多くのサービス利用者が複数のユーザ ID を保有すると言われている[2]．よって，フォーラムの電子掲示板ほど高確率に，投稿文章に付加されるユーザ ID と文章投稿者とが 1 対 1 で対応するサービスは存在しない．つまり，フォーラムの電子掲示板に投稿される文章には，文章投稿者を一意に決定する情報が存在し，この文章を用いることで，著者推定の精度評価実験が行える．なぜならば，2 節で述べる著者推定タスクの手順 4 の文体類似度順位を求めるためには，各文章の投稿者が一意に決定される情報が必要だからである．具体的には，テストデータ中の各文章と同一著者である学習データ中の文章をユーザ ID によって取得することで，学習データ中の当該文章の文体類似度順位を求める．

5.1.2. 文体類似度順位評価実験

文体類似度順位評価実験は，2 節で述べた著者推定タスクに基づいた以下の流れで行う．本評価実験により，複数の文体類似度順位を著者推定手法ごとに得る．

[1] Wikipedia ニフティサーブ http://ja.wikipedia.org/wiki/ニフティサーブ (accessed on 2013/01/07)

[2] ICT 総研「SNS 利用動向・広告活動に関する調査」http://www.cross-shop.jp/user_data/pdf/D-1201-IC-1782-P1.pdf(accessed on 2013/01/07)

表 1　5.1.2 のので評価対象となる著者推定手法

手法名	文体定量化方法	文体類似度算出方法
松浦らの手法	文字 3-gram 頻度分布	式(1):Tankard の文章相違度計算手法
中島らの手法	品詞 4-gram 頻度分布	式(2):ピアソンの積率相関係数
安形らの手法	Zip 圧縮	式(3):Zip 圧縮前後のデータ長差異
提案手法	品詞タグ・文字混合 2-gram 頻度分布	式(2):ピアソンの積率相関係数
多手法併用手法	式(7):安形らの手法，松浦らの手法，提案手法の併用	

手順 1) 学習データとテストデータの作成

次の手順で学習データとテストデータを作成する．

① フォーラムを一つ選択する．

② ①のフォーラムの電子掲示板に投稿しているユーザ ID をランダムに一つ選択する．

③ ①のフォーラムに②のユーザ ID から投稿された 20 記事を抽出し，各々の記事の冒頭 500 文字を結合し，10000 文字からなる文章を作成し，学習データとする．

④ ③と同様に，①のフォーラムに②のユーザ ID から投稿された 20 記事（ただし③で用いた記事とは別の記事）を抽出し，各々の記事の冒頭 500 文字を結合し，10000 文字からなる文章を作成し，テストデータとする．

⑤ ②においてユーザ ID を変更し，合計 N 個のユーザ ID に対して③と④を適用し，学習データ N 個，テストデータ N 個を作成する．学習データ及びテストデータ中の各文章には，対応するユーザ ID を付与しておく．

上記の手順で行われるのは，同一話題の文章に対する著者推定タスクである．これは，生成される学習データとテストデータにおいて，各々から抽出される同一著者の 2 文章は，話題が同じである同一フォーラムの文章だからである．相違話題の文章に対する著者推定タスクの場合は，手順④において記事抽出先となるフォーラムを①とは別のフォーラムにする．これにより，生成される学習データとテストデータにおいて，各々から抽出される同一著者の 2 文章は，相違話題であることが保証される．

手順 2)文体相違度算出

手順 1 で作成した学習データとテストデータの組について，著者推定タスクにおける手順 2 と手順 3 の方法で文体相違度を算出する．文体相違度算出には，表 1 で示す 5 つの手法を用いる．なお，算出する 5 つの文体相違度は別々に保持しておく．なお，各手法における n-gram の n の値を表 1 の通りにしたのは，事前実験において，上記の場合がもっとも精度が高くなったためである．また，表 1 における多手法併用手法とは，4.2 で述べた複数の著者推定手法を併用する手法のことである．当該手法では，安形らの手法，松浦らの手法，提案手法を 4.2 で述べた方法で併用した．また，当該 3 つの手法に対して設定される式(7)のa_iの値は，1.0，2.0，3.5 とした．当該多手法併用手法を採用したのは，事前実験において当該 3 手法及び各a_iの組み合わせが最も精度が高くなったためである．

手順 3)文体類似度順位算出

テストデータ中のすべての文章に対して文体類似

度順位を算出する．なお，文体類似度順位は手順 2 における手法ごとに求められる．これに加え，本実験では安形らの手法，松浦らの手法及び提案手法との最良併用時の文体類似度順位も求める．

最良併用とは，当該 3 手法を最良に併用して文体類似度順位を求めるものである．最良併用によって得られる著者推定精度は，当該 3 手法を用いる多手法併用手法による著者推定精度の理論最大性能となる．最良併用による文体類似度順位は，学習データ中の各文章について，各手法の何れかで著者推定失敗要因文章とならなければ，最良併用でも著者推定失敗要因文章とならないことに基づいて求められる．著者推定失敗要因文章とは，推定対象文章との文体相違度が，正解文章と推定対象文章との文体相違度よりも高い学習データ中の文章のことである．正解文章とは，その著者が推定対象文章の著者と一致する学習データ中の文章のことである．文体類似度順位は各著者推定手法における著者推定失敗要因文章数に 1 を加えた数になるので，最良併用における当該順位は当該 3 手法に共通する著者推定失敗要因文章数に 1 を加えた数から求められる．

5.1.3. 推定処理時間測定実験

推定処理時間測定実験では，1 つの推定対象文章と N の文章を含む学習データを著者推定タスクに与えた時，テストデータの文章の著者を推定するまでにかかった時間を測定する．当実験は，表 2 に記されるスペックを持つ計算機によって行った．実験内容は 5.1.2 の文体類似度順位評価実験における手順 2 と手順 3 を，各実験対象手法を用いて行う．実験対象手法とは，表 1 における，松浦らの手法，中島らの手法，安形らの手法，提案手法の 4 手法のことである．当実験は，文体類似度順位評価実験の手順 1 における N を多様に変化させ，各々手順 3 を処理する時間を測定することで行なう．

5.2. 実験結果の評価

5.2.1. 文体類似度順位の評価

5.1.2 の評価実験をその手順 1 において N=7,260 として行い，得られるテストデータ中の文章ごとの文体類似度順位に対して評価を行う．また本評価は，評価実験の手順 1 において，同一話題に対する文章と相違話題に対する文章の各々に対して行うものとする．なお，手順 1 で選択される 7,620 個のユーザ ID は，同一話題の文章に対する評価実験と相違話題の文章に対する評価実験で共通である．

評価は，表 1 の著者推定手法ごとに得られる複数の文体類似度順位について以下を求めることで行う．

1. Mean Top-k Recall
2. PRECISION@1
3. MRR (Mean Reciprocal Rank)

Mean Tok-k Recall は，テストデータ中の各文章に対して得られる複数の文体類似度順位について，1 位から 7,620 位の累積相対度数を得ることで求める．当評価結果は，各順位の累積相対度数分布で示される．

表 2　5.1.3 で用いた計算機のスペック

項目	性能
CPU	Intel Core i7 930 @ 2.8GHz
メモリ	DDR3 28GB
HDD	4TB Serial ATA RAID0
OS	Windows 7 Professional 64bit
Java 仮想マシン	java 1.6.0_21

Mean Top-k Recall は，著者推定タスクによって候補者群から任意の数 k 人の候補者を抽出するとき，抽出した候補者群中に実際の著者がいる確率を示す．例えば，当評価で得られた累積相対度数分布において，順位 k の累積相対度数が 0.5 なら，抽出した k 人の候補者群に 50%の確率で実際の著者が含まれることが示せる．

PRECISION@1 は，著者推定タスクで扱ったテストデータの文章群の中で，文体類似度順位が 1 位となる割合である．PRECISION@1 が高くなる手法は，高く評価される．当該評価は，2.3 で述べた既存の著者推定研究における手法評価に該当する．

MMR とは，文体類似度順位の平均逆順位で，5.1.2 の評価実験で得られる文体類似度順位の集合 R を用いた以下の式で得られる．

$$MMR = \frac{1}{|R|}\sum_{r \in R}\frac{1}{r} \tag{10}$$

MMR は，文体類似度順位の累積相対度数分布を定量的に評価したものである．具体的には，すべてのテストデータにおいて文体類似度順位が高くなるときに，MMR の値も高くなる．よって，MMR が高くなる手法は高く評価される．

当節における評価方法による結果は図 2，図 3，図 4，図 5 及び表 3 の通りである．同一話題文章に対する著者推定タスクにおける文体類似度順位の累積相対度数分布については，図 2 及び図 3 で示す．図 2 はすべての順位に対する累積相対度数分布を示し，図 3 は 1 位から 50 位における上位の順位における累積度数分布を示す．図 4 及び図 5 は，相違話題文章に対する著者推定タスクであること以外において図 2 及び図 3 と同じ意味を持つ．表 3 は同一話題文章及び相違話題文章に対する著者推定タスクの評価結果から得られ

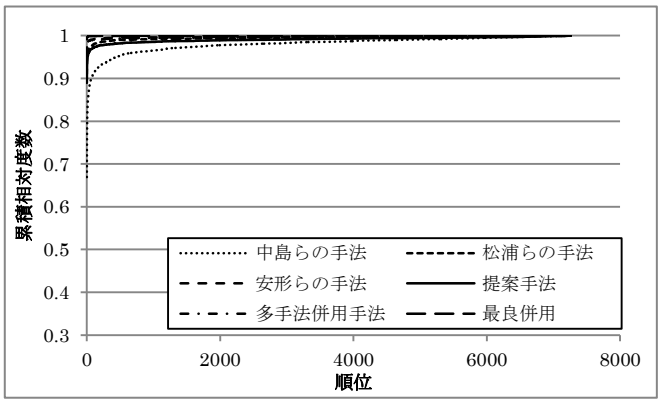

図 2　同一話題文章に対する著者推定タスクにおける全順位の累積相対度数分布

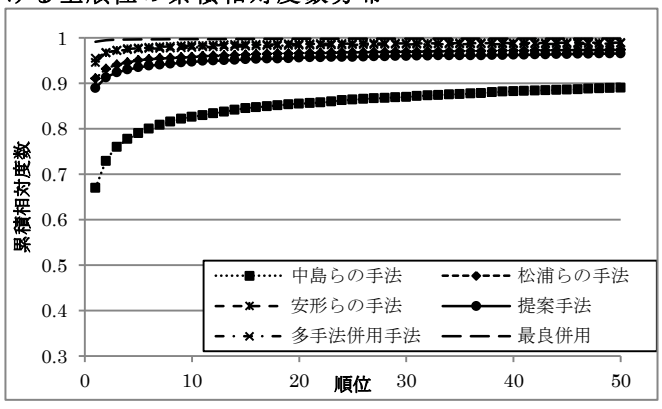

図 3　同一話題文章に対する著者推定タスクにおける 1 位から 50 位までの累積相対度数分布

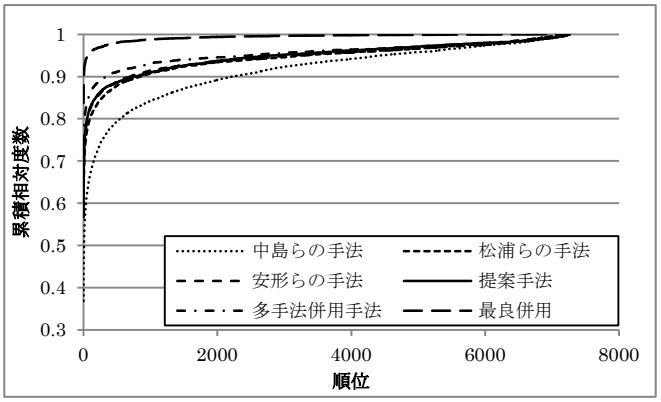

図 4　相違話題文章に対する著者推定タスク結果における全順位の累積相対度数分布

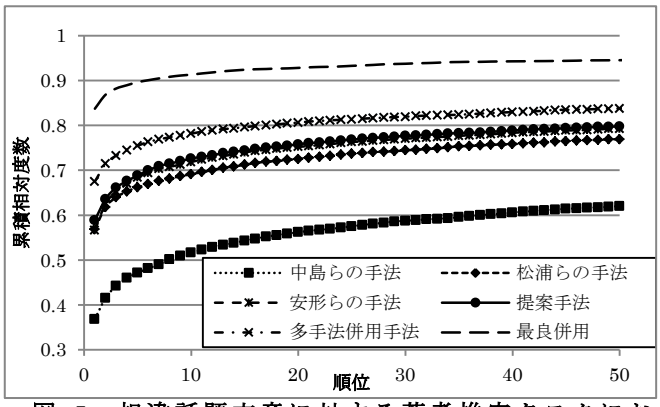

図 5　相違話題文章に対する著者推定タスクにおける 1 位から 50 位までの累積相対度数分布

た PRECISON@1 と MMR を示している．

まず，図 2，図 3，図 4，図 5 の 4 つグラフから，多手法併用手法が最も精度が高いことが分る．次いで安形らの手法が高精度となるが，相違話題文章に対する著者推定タスクでは，提案手法とほぼ同程度の精度となることがわかる．提案手法については，当該の 4 つのグラフから，中島らの手法に対し精度が向上し，改良が成功していることがわかる．上記の結果は表 3 の各評価値からも同様の評価結果を得ることができる．また，評価結果から，多手法併用手法を用いると，同一話題文章に対する著者推定タスクにおいて，10 人の候補者を抽出したとき約 97.9%の精度で実際の著者が含まれることがわかる．ただ，相違話題文章に対する著者推定タスクについては，80%の確率で実際の著者が含まれる候補者群を集めるために，多手法併用手法を用いて 17 人の候補者を抽出する必要があることがわかる．これは，安形らの手法を用いた場合における 59 人に対して，3 分の 1 の人数である．最後に，評価結果である各図を見ると,併用手法は最良併用に比べ，その精度に差があることがわかる．つまり，多手法併用手法を改良することで，さらに精度向上が期待できると言える．

5.2.2. 話題変化の頑健性評価

当評価は，著者推定タスクで同一話題の文章と相違話題の文章を用いるとき，双方の間で推定精度がどの程度変化するのかを評価する．当評価結果によって，話題変化に対する頑健性を手法ごとに評価できる．

当評価は，同一話題文章に対する 5.1.2 の評価実験で得られる著者推定手法の評価値P_Sと，同一話題文章に対する 5.1.2 の評価実験で得られる著者推定手法の評価値P_Dを用いて，以下の式(11)の頑健性$Robustness$を

表 3　5.2.1 の評価結果における各種評価値

手法	推定対象	同一話題文章		相違話題文章	
	評価値	PRECISON@1	MMR	PRECISON@1	MMR
中島らの手法		0.72	0.67	0.42	0.37
松浦らの手法		0.93	0.91	0.61	0.57
安形らの手法		0.96	0.95	0.62	0.57
提案手法		0.91	0.89	0.64	0.59
多手法併用手法		0.96	0.95	0.71	0.67
最良併用		0.99	0.99	0.86	0.84

表 4　5.2.2 における話題変化の頑健性評価結果

手法	評価値	PRECISON@1	MMR
中島らの手法		0.58	0.55
松浦らの手法		0.66	0.62
安形らの手法		0.65	0.60
提案手法		0.70	0.66
多手法併用手法		0.74	0.71
最良併用		0.87	0.84

算出し，当評価を行う．$Robustness$は，評価値P_Sに対する評価値P_Dの割合を算出することで，同一話題文章を用いて得られる評価値P_Sが，相違話題文章を用いることでどれほど評価値P_Dが低下するかを表現できる．

$$Robustness = \mathrm{Min}\left(1, \frac{P_D}{P_S}\right) \tag{11}$$

当評価で用いる評価実験の評価値には，表 3 における PRECISON@1 及び MMR を用いる．当評価結果は，表 4 に示す．

表 4 から，提案手法及び多手法併用手法の PRECISON@1 及び MMR の評価値による頑健性は，既存手法よりも高いことがわかる．この結果は，提案手法や多手法併用手法が話題変化に頑健な手法であることを示し，同一話題文章収集困難化の問題に影響されにくいことがわかる．

5.2.3. 推定処理時間評価

当評価は，5.1.3 の評価実験を行った際に得られる，手法ごとの処理時間に基づいて，各手法の著者推定処理時間を評価する．当評価では，5.1.2 の評価実験の手順 1 の N を 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000 としたときの各々の結果について評価する．ただし，5.1.3 の評価実験を 10 回行い，各々得られる処理時間の平均値を算出したものを，当評価で用いる処理時間とする．

評価結果である図 6 より，提案手法は松浦らの手法及び安形らの手法よりも高速に処理ができることがわかる．また，学習データ中の文章数を表す N が大きくなるに連れて，提案手法と松浦らの手法及び安形らの手法の間で著者推定処理時間に差が出ることがわかる．これは,学習データ中の文章数は候補者数を表すため，より大規模な候補者群を扱うと，松浦らの処理及び安形らの手法の遅さが顕著となることを示している．なお，表 1 における多手法併用手法は，安形らの手法，松浦らの手法及び提案手法の著者推定処理結果を使うため，当該 3 手法の推定処理時間の和より多くの推定処理時間が必要となる．

5.3. 評価結果の考察

本節で行った評価実験から，提案手法と多手法併用手法について既存手法とは異なる特徴を持つことがわかった．提案手法による著者推定は，安形らの手法よ

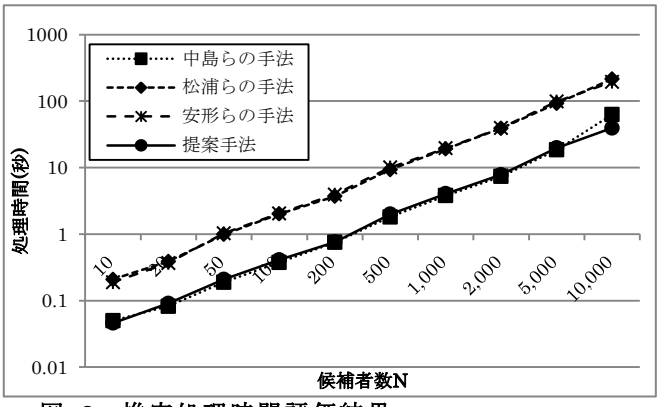

図 6 推定処理時間評価結果

りも高速に推定処理ができるが，安形らの手法と同等の推定精度となる．多手法併用手法による著者推定はどの既存手法よりも高い推定精度となるが，その処理時間は比較的長い．ただし，提案手法及び多手法併用手法における著者推定で用いる文章中の話題に対する頑健性は，どの既存手法よりも高い．

上記の結果は，大規模候補者群に対する著者推定において，著者推定を行う状況に応じて手法を使い分けられることを示している．具体的には，文体類似著者の高頻出の問題に伴う推定精度低下を重点的に対応するのであれば，最も高精度に著者推定可能な多手法併用手法を用いることが望ましい．また，推定処理時の計算量増加に伴う処理速度低下を重点的に対応するのであれば，処理時間が短い提案手法を用いることが望ましい．なお，提案手法及び多手法併用手法の双方とも話題変化に対する頑健性の高いことから，双方とも同一話題文章収集困難化の問題に対応することができている．

6. まとめ

本稿では，既存の著者推定で取り扱って来なかった，大規模候補者群に対する著者推定について，推定手法と手法評価方法の提案を行った．本稿で提案した著者推定手法を用いることで，大規模候補者群に対する著者推定において，高精度または高速な推定が行えることがわかった．これは，大規模候補者群に対する著者推定で顕著化する「文体類似著者の高頻出」「同一話題文章収集困難化」「推定処理時の計算量増加」の 3 つの問題に，提案手法が各々対応できるためである．

本研究の課題点として，相違話題の文章に対する著者推定精度が低いことが挙げられる．具体的には，7260 人の候補者群から 90%の確率で実際の著者を含む候補者群を抽出するとき，その候補者数が 340 人となるのは，実用的と言い難い．多手法併用手法と最良併用の間で推定精度の差があることから，より高精度に推定可能な多手法併用手法が考案可能であり，今後の研究で実現していくことが求められる．

謝 辞

本研究は，ニフティ株式会社様より提供していただいた，ニフティサーブのフォーラムにおける電子掲示板のデータによって実現した研究である．データを提供していただいたニフティ株式会社様に，謹んで感謝の意を表する．

参 考 文 献

[1] 松浦司，金田康正: “近代日本文学者 8 人による文章における文字 n-gram の分布を利用した近代日本語文の著者推定”，計量国語学， Vol.22, No.6, pp.1-9, 2000.

[2] 安形輝 :“圧縮プログラムを応用した著者推定”，J. of Library and Information Science, 三田図書館・情報学会, No.54, pp.1-18, 2005.

[3] 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.

[4] 石川尚季，西村涼，渡辺靖彦，村田真樹，岡田至弘: “コミュニケーションサイトに投稿されたメッセージに対する著者の推定”，信学技報 (NLC) ,Vol.109, No.142, pp.79-84, 2009

[5] 佐藤進也，原田昌紀，風間一洋: “文字列出現頻度比較による情報源間の類似性判定”，情処研報 (DD), Vol.2002, No.28, pp.119-126, 2002

[6] 中島泰，山名早人: “品詞と助詞の出現パターンを用いた類似著者の推定とコミュニティ抽出”，DEIM2011, B6-5, 2011.

[7] 坪井祐太，松本裕治: “異なるタイプのドキュメントに対する著者推定”，情処研報(NL)，Vol.2002, No.20, pp.17-24, 2002.

[8] 井上雅翔，山名早人: “品詞 n-gram を用いた著者推定手法 : 話題に対する頑健性の評価”，日本データベース学会論文誌，Vol.10，No.3，pp.7-12, 2012.

[9] 田代光輝，鈴木隆一，松井くにお，宇田周平，折田明子，三浦麻子，森尾博昭: “NIFTY-Serve フォーラムの全データの整形”，第 5 回知識共有コミュニティワークショップ, 2012

[10] フリードマン，リチャード・エリオット 著，松本英昭 訳: “旧約聖書を推理する : 本当は誰が書いたのか”，海青社, p.355, 1989

[11] 村上征勝，“著者を探る古文書の計量分析”，信学誌, Vol.85, No.3, pp.158-161, 2002

[12] 細江光: “谷崎の作品ではなかった 偽作「誘惑女神」をめぐって”，国文学 解釈と教材の研究，学灯社, Vol.33, No.8, pp.134-137, 1988.

[13] Stamatatos, E.: “A Survey of Modern Authorship Attribution Methods”, J. of the American Society for Information Science and Technology, Vol.60，No.3, pp.538-556, 2009.

[14] N.V. Chawla, N. Japkowicz and A. Kotcz : “Editorial: special issue on learning from imbalanced data sets”, J. of the ACM SIGKDD Explorations Newsletter, Vol.6, No.1, pp.1-6, 2004.

[15] J. Tankard: “The Literary Detective”, BYTE, February, No.2, pp.224-227, 1986

[16] 形態素解析システム Sen, http://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html (accessed on 2013/01/06)