RF-003

性別による言葉遣いの違いを考慮したブロガーの年齢推定手法の提案

Suggestion of the Age Estimate Technique of the Blogger in Consideration of Difference in Gender-related Language

古山 直樹† 筧 捷彦‡
Naoki Furuyama Katsuhiko Kakehi
†早稲田大学大学院基幹理工学研究科情報理工学専攻 〒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
E-mail: †furuyamanaoki@suou.waseda.jp, ‡kakehi@waseda.jp

1. はじめに

近年，インターネットの普及によって，一般の人々の意見がインターネット掲示板やブログ，ツイッターといった様々な Web コンテンツを通して発信されるようになった．これらの意見は，マーケティングや意識調査などにおいて有用である．

一方，Web 上の情報は匿名性が高いため，情報の発信者の人物像が分かりにくい．もし，情報の発信者の人物像が分かれば，それを基により詳細な調査を行うことができるだろう．例えば，自社の若者向け化粧品の評判を知りたいといった場合，人物像が分かっていれば，“20 代の女性”といったように人を限定して評判を知ることができる．

本研究では，一般の人々の意見が記述される Web コンテンツとして，ブログを選択した．ブログを選択した理由は，情報量が多く，リアルタイム性に優れ，また，個人のホームページなどに比べて html の構造が単純なために解析が行いやすいからである．

ブログから推定する情報として，その書き手であるブロガーの性別・年齢に着目した．性別や年齢は，個人の趣向や考え方に大きく関与すると考えたからである．推定を行うために，既存の研究やシステムを調査し，それらよりも高い精度で推定ができる手法を考案する．

2. 関連研究

ブロガーの性別や年齢を推定する研究は既に行われている．ここでは，それらの研究について紹介していく．

性別推定[1]では，男女間での話し言葉の性質の違いを考慮した素性を用い，Support Vector Machine(SVM)による手法によって高い精度で推定が行われた．年代推定[2]では，年齢を 10 代，20 代，30 代，その他という 4 つのクラスに分類し，情報エントロピーを用いてクラス毎の特徴語を抽出し，ブロガーがどのクラスに属するかを推定した．年齢推定[3]では，1 才毎という年代よりもさらに細かい区分で推定を行うために，ブースティングに基づく手法を用いた．英語のブログ記事を対象とした研究[4]では，ブログのスタイルや単語の特徴を素性として，Multi-Class Real Winnow を分類器に用いてブロガーの性別・年代を推定した。

ブログから情報を抽出するシステムには，ブログ通信簿[5]がある．このシステムでは，ユーザにブログの URL を入力させ，そのブログの本文や RSS に対して解析を行う．そして，特徴的な単語や投稿日時から，ブログの性別や年齢，更新頻度や興味のある分野といった情報を抽出・推定し，出力する．

3. 年齢推定方法の提案

本研究では，ブロガーの年齢を推定することに焦点を置く．既存の年齢推定手法に，性別による語の使い方の違いを考慮した手法を加え，より高い精度で推定を行うことを目指す．具体的には，ベイジアンフィルタを利用した性別推定手法を考案し，SVM による手法との性能比較を行う．そして，ブースティングに基づく年齢推定手法に性別推定の結果を考慮するという提案手法で実験を行い，性能比較を行う．

3.1 ブロガーの性別推定

ブログ記事の本文部分から，ブロガーの性別推定を行う方法について説明する．

文章の書き方は，男性と女性で傾向が異なると考えた．例として，一人称代名詞の使い方を挙げてみる．男性は“俺”や“僕”といった語を多く使い，女性は“私”や“あたし”といった語を多く使う傾向があるだろう．このような性別による語の使い方の違いは，性別推定に利用できると考えた．意味を持つ最小の文字列である形態素を推定の素性に用いて，性別推定を行う．

判定のための道具として，Paul Graham の方式によるベイジアンフィルタ[6] を用いる．これは，ベイズの定理を簡略化したものに基づいており，スパムメールの検出等に利用されている．メールをスパム・ハムの二種類に分類する，という操作をブログの場合に置き換え，ブログを男性が書いたブログ・女性が書いたブログの二種類に分類する．

ブロガーの性別推定を行うには，推定を行うシステムに予め必要な情報を学習させておく必要がある．そのための準備として，男性・女性の各ブログ記事で使われる形態素の傾向を書き込んだ辞書データを作成する．

学習データとなるブログ記事が与えられたとき，そこに含まれる形態素 w が，1 記事あたり平均何回出現するかを性別毎に求める．形態素 w の，男性のブログ記事での平均出現回数 $A_M(w)$ と女性のブログ記事での平均出現回数 $A_F(w)$は，それぞれ式(1)で表される．

$$A_M(w)=\frac{C_M(w)}{N_M},\quad A_F(w)=\frac{C_F(w)}{N_F} \tag{1}$$

ただし，$C_M(w)$・$C_F(w)$は男性・女性のブログ記事での形態素 w の総出現回数，N_M ・N_F は男性・女性のブログ記事の総数をそれぞれ表している．

これらの値を用いて，形態素 w の男性らしさ $P_M(w)$を，式(2)で定義する．

$$P_M(w)=\frac{A_M(w)}{A_M(w)+A_F(w)} \tag{2}$$

$P_M(w)$が 1 に近ければ男性のブログ，0 に近ければ女性のブログで，形態素 w が使われる傾向が強いということになる．学習データに現れる全ての形態素について男性らしさを求めたものが，辞書データとなる．男性・女性の各ブログで使われる傾向が強かった形態素の例を表 1 に示す.

表 1:男性・女性のブログで使われやすかった形態素の例

形態素	男性らしさ	形態素	男性らしさ
鉄道	0.869	夫	0.087
バイク	0.816	洗濯	0.236
阪神	0.761	素敵	0.197
として	0.657	ちゃん	0.249
僕	0.753	あたし	0.101

作成した辞書データを用いて，性別の判定を行う．性別を知りたいブログ記事 B に現れていて，さらに辞書データにも登録されている形態素のうち，男性らしさの値が 0.5 から離れている上位 k 個の形態素の集合を$[B]_k$ と表すことにする．これは，男性または女性のブログ記事で使われる傾向が強い形態素の集合を意味している.

男性らしさ $P_M(w)$と形態素の集合$[B]_k$ を用いて，ブログ記事 B の著者が男性である確率 $P_k(M|B)$を式(3)から求める.

$$P_k(M \mid B) = \frac{\prod_{w \in [B]_k} P_M(w)}{\prod_{w \in [B]_k} P_M(w) + \prod_{w \in [B]_k} (1 - P_M(w))} \tag{3}$$

本研究では表 2 のように閾値を定め，性別の判定を行うことにした.

表 2:性別判定の閾値の設定

男性である確率 $P_k(M\|B)$	判定
$P_k(M\|B) > 0.9$	男性
$0.9 \geqq P_k(M\|B) \geqq 0.1$	性別不明
$P_k(M\|B) < 0.1$	女性

3.2 ブロガーの年齢推定

ブログ記事の本文部分から，ブロガーの年齢推定を行う方法について説明する.

文章の書き方は，性別のときと同様に年齢によっても傾向が異なると考えた．例として，使用する名詞について考える．10 代のブロガーには学生が多いため，“学校” や “テスト” といった語が多く使われるだろう．一方，30 代のブロガーには主婦や社会人が多いため，“洗濯” や “仕事” といった語が多く使われるだろう．このような，年齢間での語の使い方の差に着目し，年齢の推定を行う.

年齢推定は，性別推定に比べて判定するクラス数が多いため，性別推定で用いた手法では推定が難しいと考えた．そこで，年齢推定では，ブースティングに基づく手法を用いることにした.

ブースティングに基づく手法[3]では，年齢推定に用いる素性として付近に現れる共起語を使用している．本研究でも同様の素性・手法を用いることにした．素性毎に重み付き分類器を作成し，それらを統合して最終的な分類器とし，出力値が最大となる年齢を推定年齢とする．年齢 x における最終的な分類器 $V(x)$は式(4)で定義される.

$$V(x) = \sum_f w_f g_f(x) \tag{4}$$

$$g_f(x) = \frac{\text{年齢}x\text{で素性}f\text{が使われた記事の割合}}{\text{各年齢で素性}f\text{が使われた記事の割合の合計}}$$

$$w_f = P * \log((\log_2 n - E_f) * Q + 1)$$

素性 f の分類器は $w_f g_f(x)$で定義される．ここで，$g_f(x)$は年齢 x における素性 f の利用率に基づく推定量を表している．w_f は素性 f の年齢の特定しやすさを意味する重み付けを，情報エントロピーを用いて式化している．P,Q は重み付け行う際のパラメータ値であり，n は推定候補となる年齢数である．推定量の分布の偏り E_fは，情報エントロピーを用いて式(5)で表される.

$$E_f = \sum_x \left(-\frac{g_f(x)}{T_f} \log_2 \frac{g_f(x)}{T_f}\right) \tag{5}$$

$$T_f = \sum_x g_f(x)$$

例えば，“学校”+“給食”という素性があった場合，その素性がよく使われるであろう 10 代前半の年齢帯では推定量の値は大きくなり，それ以外の年齢帯では小さくなる．また，一定の年齢だけで多く使われる素性は，年齢を特定するのに有効であるため，重み付けは大きくなり，結果的に分類器の出力値も大きくなる．逆に，どの年齢でも同程度に使われる素性は，重み付けは小さくなるため，分類器の出力値も小さくなる．個々の素性からは正確な年齢が分からなくても，それらを総合して推定することで，有効な結果を得ることができるのである.

本研究では上述の手法に加え，先に行った性別推定の結果も推定に組み込むことにした．これは，同じ年齢でも男女間で語の使い方には差があり，それを利用することでより正確な推定ができるという考えに基づく.

提案手法について説明する．まず，学習データとなるブログ記事を次の 3 種類に分ける.

A :年齢が公開されている男性のブログ記事
B :年齢が公開されている女性のブログ記事
C :年齢が公開されているが性別が分からないブログ記事

A だけを使って作った男性用の分類器，B だけを使って作った女性用の分類器，A・B・C を使って作った性別不明用の分類器を作る．判定の際には，初めに性別推定を行い，その結果に応じた分類器で年齢の判定を行う．このようにすることで，既存の手法よりも高い推定精度が出せると考えた.

4. ブロガーの情報推定システムの設計

提案手法の評価を行うために，プロトタイプシステムを作成した．本システムでは，ブロガーの性別と年齢の推定を行う．システム全体のイメージを図 1 に示す.

ブロガーの性別推定では，ブログ記事の本文からブロガーの性別を男性・女性・性別不明の 3 種類に判定する．ブロガーの年齢推定では，ブログ記事の本文と性別推定の出力結果から，ブロガーの年齢を 1 才単位で判定する.

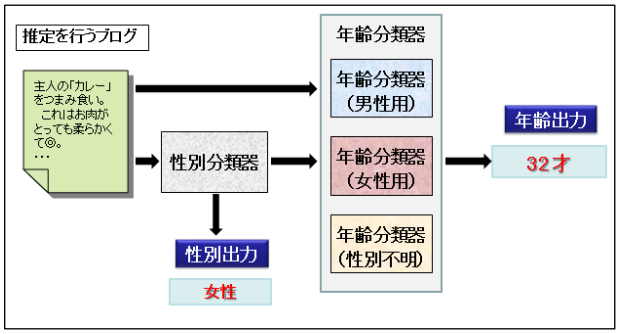

図 1:ブロガーの情報推定システム

ブロガーの情報を推定するためには，推定の指標となる辞書が必要となる．この辞書は，収集したブログデータを基にして作成される．

ブログサービスの中には，性別や年齢といった，ブロガーのプロフィールを記述する欄が設けられているものがある．この欄に記述されている性別と実際の性別は，95 %以上一致すると報告されている[1]．この結果から，ブロガーのプロフィールを記述する欄，すなわちプロフィール欄に書かれている情報は，実際のものと一致すると考えた．プロフィール欄に情報が書かれているブロガーのブログを収集し，学習データやテストデータとして使用する．

本研究では，プロフィール欄が設けられているブログサービスの中から Yahoo!ブログ[7] を選択した．Yahoo! ブログで作られたブログをランダムに表示させる，“ランダムブログ” 機能を用いて，ブログデータの収集を行った．

収集したブログデータのうち，ブログ記事の本文部分は文章であるため，推定を行うために形態素毎に切り分ける必要がある．そこで，形態素解析エンジン Mecab[8]を用いて文章の切り分けを行った．

ブログデータの収集は，2010 年 9 月から 2 ヵ月に渡って行った．Yahoo!ブログでは，プロフィールの公開は必須ではない．そのため，収集したブログデータの中には，性別や年齢が片方だけしか公開されていないブログや，どちらも公開されていないブログも含まれていた．収集したブログ 42326 アカウント分について，性別と年齢の公開状況を表 3，表 4 に示す．また，年齢の分布状況を図 2 に示す．

表 3:性別公開状況

	アカウント数
男性	17045
女性	13576
非公開	11705

表 4:年齢公開状況

	アカウント数
公開	3987
非公開	38339

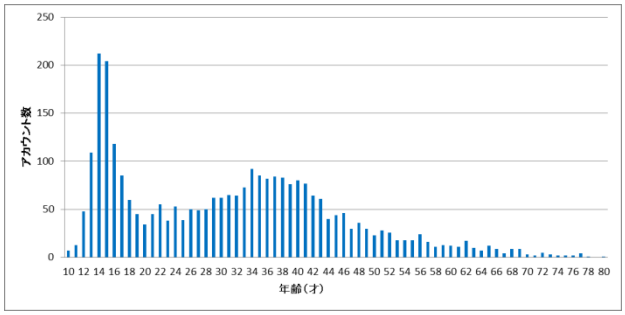

図 2:年齢が公開されているブログの年齢分布

5. 評価実験

本研究の手法の有効性を示すために，性別推定と年齢推定の性能を評価する実験を行った．

5.1 性別推定

性別推定の性能を評価するための指標には，精度(precision)，再現率(recall)，F 値(F-measure)を用いた．

精度は，判定された結果がどれだけ正しいかという割合を表す．再現率は，判定対象からどれだけ正解を得ることができたかという割合を表す．

精度と再現率はトレードオフの関係にある．そこで，両方の調和平均をとる F 値を性能評価に用いた．F 値は 0 以上 1 以下の値をとり，値が大きいほど性能が良いことを意味する．

$$\text{精度} = \frac{\text{正解数}}{\text{性別が判定された数}}$$

$$\text{再現率} = \frac{\text{正解数}}{\text{性別が判定された数} + \text{性別不明の数}}$$

$$\text{F値} = \frac{2 \cdot \text{精度} \cdot \text{再現率}}{\text{精度} + \text{再現率}}$$

収集したブログデータのうち，男性のブログ記事 50000 件と女性のブログ記事 50000 件をランダムに選択し，実験に使用した．辞書データを作るための学習データとテストデータの分割方法には 4 分割交差検定法を用い，結果の平均値を実験結果の値とした．

実験は，男性のブログを“男性” と判定することができるか，女性のブログを“女性” と判定することができるか，と男女で分けて行った．使用形態素数 k を変えると，精度，再現率，F 値がどう変化するかを調べた．結果を図 3，図 4 に示す．

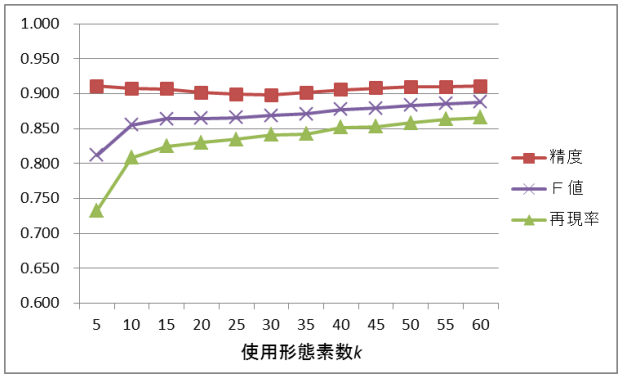

図 3:男性のブログの性別推定結果

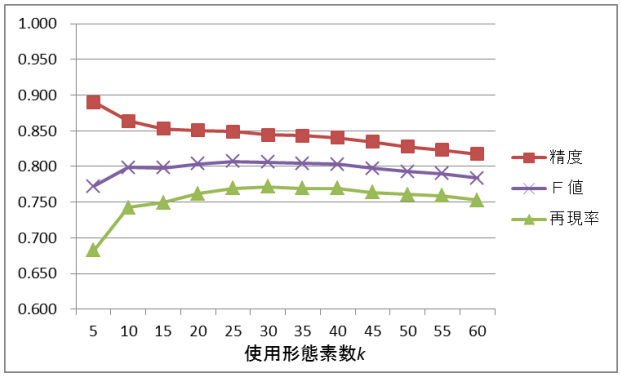

図 4:女性のブログの性別推定結果

実際の推定では，判定を行いたいブログの性別は分からないので，最適な k の値を決める必要がある．最適な k を決める指標として，男性と女性のブログでの F 値の平均が最大のものを選ぶことにした．その結果，k = 40 のときに F 値の平均が最大値 0.84 となったため，本システムでは k=40 として推定を行うことにした．

既存研究の手法[1]では，同様の実験を行い，男性のブログでは精度 0.91，再現率 0.79 を女性のブログでは精度 0.95，再現率 0.81 という結果が出ている．この結果の F 値は 0.86 なので，本研究の手法でも同等の性能で推定が行えていると言える．

5.2 年齢推定

年齢推定では，判定するクラス数が年齢の数だけ存在するため，正確な年齢を推定することは難しい．そこで，本研究では既存研究[3]と同様に，推定された年齢に許容誤差範囲を設けて，正解の年齢がその範囲内にあれば正解，そうでなければ不正解として精度を求めることにした．許容誤差範囲を変えると精度がどう変わるかを調べることで，推定された年齢がその範囲内でどの程度の信頼性があるのかが分かる．

図 2 で示したように，収集したブログデータの年齢分布には偏りがある．収集数が少ない年齢では，十分な量の学習を行うことができない．また，ブログ記事に含まれる形態素数が極端に少ないブログも，同様の理由で適さない．そこで，本研究では 12 才から 50 才までのブログで 50 形態素以上を含むブログ記事のみを，学習データやテストデータとして使用することにした．

収集したブログデータで上記の条件を満たすもののうち，24316 件をランダムに選択し，18237 件の記事を学習データに，6079 件の記事をテストデータに使用した．重み付けを行う際のパラメータ値は P=Q=10 とし，既存研究の手法と，既存研究の手法に性別推定の結果を組み込むという提案手法の 2 通りで実験を行い，結果を比較した．結果を図 5 に示す．

結果を比較してみると，本研究の手法の方が平均 5.4% 高かった．よって，同じ年齢でも性別の男女間で語の使い方には差があり，性別推定の結果を年齢推定に組み込むという提案手法は有効であることが示せた．

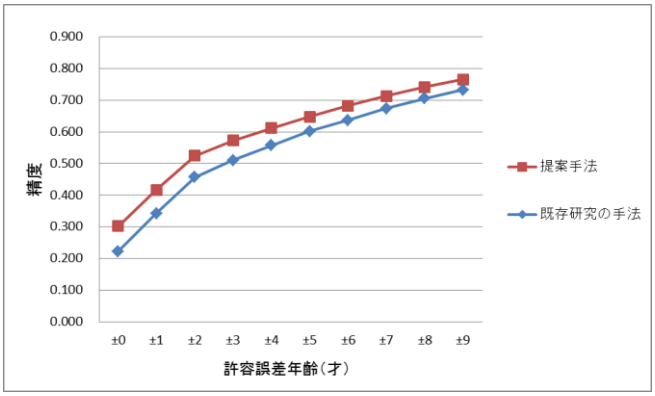

図 5:既存研究の手法と提案手法での年齢推定結果

6. まとめ

本研究では，ブロガーの性別と年齢を推定するための手法を考案し，制作したシステムによって評価を行った．

性別推定では，ベイジアンフィルタを利用した推定手法を提案し，既存研究と同程度の性能で推定を行うことができた．

年齢推定では，ブースティングに基づく手法に性別推定の結果を考慮するという手法を提案した．性別推定の結果を考慮していない従来の手法との比較実験を行った結果，精度の向上を確認することができた．よって，性別による言葉遣いの違いを考慮することは，年齢推定に有効であることが分かった．

性別や年齢の推定では，学習データやテストデータの中に，スパムブログや企業・団体のブログ，他の文章を引用しているだけといった，本人が記述していないブログ記事が含まれていた．誤判定の原因には，それらのブログの影響が挙げられる．

そこで，本人が記述していないブログ記事を自動的に除去するような仕組みを加えることによって，更なる精度の向上が期待できる．今後は，そのようなブログ記事を自動的に除去する仕組みを考案していきたい．

参考文献

[1] 池田大介，南野朋之，奥村学，“blog の著者の性別推定”，言語処理学会第 12 回年次大会，2006

[2] 泉雅貴，三浦孝夫，塩谷勇，“Blog 著者年代推定のためのエントロピによる特徴語抽出”，DEWS，2008

[3] 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，2009

[4] J. Schlar，M. Koppel，S. Argamon and J. Pennebaker，“Effects of Age and Gender on Blogging”，AAAI，2006

[5] ブログ通信簿，goo ラボ，2011 年 4 月 15 日訪問，http://blogreport.labs.goo.ne.jp/

[6] Paul Graham，“Better Bayesian Filtering”，2011 年 4 月 15 日訪問，http://paulgraham.com/better.html/

[7] Yahoo!ブログ，YAHOO JAPAN，2011 年 4 月 15 日訪問，http://blogs.yahoo.co.jp/

[8] オープンソース形態素解析エンジン Mecab(和布蕪)，京都大学情報学研究科，日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト，2011 年 4 月 15 日訪問，http://mecab.sourceforge.net/