スケーラブルで汎用的なブログ著者属性推定手法

大倉　務†　清水　伸幸††　中川　裕志††

† 東京大学大学院 情報理工学系研究科
†† 東京大学 情報基盤センター
E-mail: †{ohkura,shimizu}@r.dl.itc.u-tokyo.ac.jp, ††nakagawa@dl.itc.u-tokyo.ac.jp

あらまし　本論文では，ブログの著者属性推定問題を扱う．ブログを用いた流行分析が広がりつつあるが，その際に年齢・性別・居住域などの著者属性が分かればその有用性はさらに高まる．これまでに，いくつかのブログの著者属性推定手法が提案されてきたが，汎用的なものではなかった．本論文では著者属性推定問題を，個々の属性固有の性質を利用しない単純な多クラス文書分類問題ととらえ，χ^2 値による素性選択と Complement Naive Bayes を用いる方法を提案する．その上で提案手法を現実のブログデータに適用する実験を行い，汎用的であるにも関わらず高速かつ高精度に著者属性を推定できることを示す．
キーワード　ブログ，著者属性，文書分類

Scalable and General Method to Estimate Blogger Profile

Tsutomu OHKURA†, Nobuyuki SHIMIZU††, and Hiroshi NAKAGAWA††

† Graduate School of Information Sience and Technology, The University of Tokyo
†† Interfaculty Initiative in Information Studies
E-mail: †{ohkura,shimizu}@r.dl.itc.u-tokyo.ac.jp, ††nakagawa@dl.itc.u-tokyo.ac.jp

Abstract　We propose a general and scalable method to estimate bloggers' unstated profiles. Recently, trend analysis based on weblogs is gaining popularity, and blogger profiles provide us more detailed interpretation of data. None of previous studies proposed a method generally applicable to different attributes. In this paper, we reduce blogger profile estimation to text classification, using Complement Naive Bayes with feature selection based on χ^2 value. We applied our proposed general method to real weblog data, and experimental results show the its effectiveness and scalability.
Key words　Weblog, Profile, Text classification

1. 背　　景

ブログは，1997 年末に誕生してからの 10 年で，手軽な情報発信の方法として急速に広まった．特に日本では，元々オンライン日記を好む国民性と相まって急速に普及し，最近では新しく生まれる記事の言語は日本語が最も多い．現在，日本語によるブログは 200 万サイトから 400 万サイト程度あるといわれ，日々50 万記事以上が新しく投稿されている．

ブログに投稿される記事はユーザーの率直な意見を反映しており，また些細な事柄も含まれているといわれている．アンケートを採るなどの人手で行う流行調査と比べ，ブログのデータの分析による調査は低コストかつ短期間で実施できるため，ブログは流行調査を行うための情報元として有望視されている．ブログの記事は電子的な形で提供されているため，これらの調査は完全に機械化することができそうに思える．しかし，現状ではまだ機械では困難として，調査したい製品名を含む記事をリストアップし人手で読み解く方法が主である．

マーケティング等で調査をする際，その調査対象者の属性が重要となる．具体的には，性別，年齢，行動圏，等々である．もしこれらの属性を機械で適切に推定できれば，ブログを用いた自動調査の有用性は飛躍的に向上することが期待される．このための研究は既にいくつか行われているが，個々の属性の特徴に依存しない汎用的なものはまだない．特定の属性にとらわれない一般的な

アルゴリズムを提案できれば，調査ごとに必要となる属性を機械的に推定することが可能になり，ブログを用いた調査の適用可能範囲を大きく広げられると考えられる．このため，本研究ではブログの記事を用いその情報源であるブログ著者の属性を推定するための汎用的なアルゴリズムを提案し，実際に性別，年齢層，居住地域の 3 種類の異なった著者属性の推定が実用的な精度で行えることを確認した．

提案手法では，ブログサイトの著者属性推定をブログサイトを単位とした分類問題と見なし，線形時間で学習可能な多クラス分類アルゴリズムである Complement Naive Bayes (CNB) を適用する．分類に使用する素性は教師データ中での χ^2 値を基準に選択する．この組み合わせの特徴は，サンプルデータ以外の情報を必要とせず，大量の教師データによる学習が実用的な時間で行えることである．実際のデータで実験した結果，高速かつ実用的な精度で著者属性の推定が行えることが確かめられた．

以下，本論文では，第 2. 章で既にこの分野で行われている研究を紹介し，その問題点を指摘する．3. 章で，提案手法を紹介する．4. 章では現実のブログデータに提案手法を適用した実験の手順を示し，結果を考察する．5. 章はまとめである．

2. 関 連 研 究

ブログの著者属性の推定というタスクに注目し，個々の属性を扱った既存研究は 2 つある．1 つは安田ら [1] によるもので，ブログ著者の居住都道府県を推定するものである．安田らはブログ著者の居住都道府県を推定する手法に関して実験を行っている．「ブログテキスト中に出現する地名を，地名辞書を用いて都道府県にマッピングし，投票を行う.」という手法 (精度 48.2%) をベースに，地名が出現する文が居住域を示すか否かを別の分類器で判断したり，コーパスから各語の地域を特徴づける強さをスコア化して利用するなどの試みを行っている．しかし，どちらの手法も精度に僅かしか貢献しなかったとの実験結果を報告している．安田らはこの結果から，地名辞書ベースの手法では限界があるという知見を得たとしている．

安田らの手法は，論文中で指摘されている通り，辞書を使っているためカバレッジが低い．また，属性に特有の要素を用いており，汎用的な手法ではない．

もう 1 つの既存研究は池田ら [2] によるもので，ブログ著者の性別を推定する手法を提案している．ブログテキスト中から特定の形態素の単語を集め，2 クラス文書分類問題として Support Vector Machine (SVM) を適用して性別を推定している．SVM 出力に閾値を設け (男性, 女性, 不明) の 3 つに分類し，86.6%のブログについて性別を推定できそのうちの 88.9%が正解だったと報告している．またこの結果を人手による性別推定と比較し，ページのデザインや画像等を考慮しなくとも，テキストの内容のみから性別を推定できることが示されている．

池田らの手法は文書分類問題に帰着させていて，また素性となる単語を辞書を用いるのではなくコーパスから自動で選択するなど，性別という属性に強く密着した手法ではない．しかし，学習に時間のかかる SVM を利用しており，これは大規模な教師データを利用する障害となる．また閾値を設ける方法は各属性値を持つデータが大量にある場合は問題ないが，そうでない場合には望ましい方法とはいえない．

3. 提 案 手 法

本論文で提案する手法は，ブログデータをマーケティング調査等に使用する際に，調査者が任意の属性を切り口にした調査を行うことを可能にする手法の構築を目指して構成した．実用性を考えれば，属性ごとに異なったアプローチをとることは難しい．また，現実的な教師データの集め方としてはサービスプロバイダの 1 つが内部で持っている情報を利用するといったことが考えられるが，この場合に利用できる教師データは最低でも数万サイト分の著者属性を利用できると思われる．

これらのことを考慮し，下記の 4 点を要件とした．

（1） 数万サイト規模の教師データによる学習が可能なこと

（2） 多クラスの属性にも適用できること

（3） 推定が高速であること

（4） 各属性固有の特性を使用せず，ブログデータと教師ラベルのみから属性を学習できること

まず，各ブログサイトの記事中に出現する単語を bag-of-words モデルを用いてモデル化する．その後，高速な多クラス文書分類アルゴリズムである CNB を用いて属性値を推定する．分類器の学習は，ラベル付きデータを用いる．また，出現する全ての単語を扱うと学習時・推定時共に時間がかかってしまうので，著者属性の推定に有用な語を χ^2 に基づいて選択することとした．

本章ではまず，χ^2 値による素性選択の方法と，高速な多クラス分類アルゴリズムである CNB を紹介する．その後，これらのアルゴリズムを用いた著者属性推定に関する提案手法を紹介する．

3.1 χ^2 値による語の選択

属性の推定に有用な語を選択するための方法として，池田ら [2] は χ^2 値による選択を用いている．本研究でも同じく χ^2 値による素性選択を用いた．χ^2 値による手法その他の素性選択の手法に関しては, [3] に詳しい．

χ^2 値とは，ある集合に対する複数の分割がどの程度一

致しているかを示す指標である．ここでは，ブログサイトに，ある単語 t が含まれてるか否かによる分割と，ブログサイトの著者のある属性値が c であるか否かの分割が，どの程度一致するかを測る．この値が大きいほど単語 t が属性推定に有用であるといえる．

ここで，A を単語 t を含み属性値が c であるブログサイトの数，B を単語 t を含み属性値が c でないブログサイトの数，C を単語 t を含まず属性値が c であるブログサイトの数，D を単語 t を含まず属性値が c でないブログサイトの数，としたとき，χ^2 値は

$$\chi^2 = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

と計算される．そこで，

$$\chi^2_{max}(t) = \max_c \chi^2(t, c)$$

を計算し，$\chi^2_{max}(t)$ の値が大きい順にいくつかの語を属性推定に用いること素性とした．利用した素性の数（語の数）と推定精度の関係については 4. 章で実験結果を示す．

3.2 Multinomial Naive Bayes

Naive Bayes 法を用いて多クラス分類を行う場合，Multinomial Naive Bayes を用いるのが最も自然である．m 個のクラスへの分類問題を考えたとき，θ_{ci} を属性値 c をもつブログサイトで単語 i を含む記事が投稿される確率とし，$\vec{\theta_c} = \{\theta_{c1}, \theta_{c2}, \cdots, \theta_{cn}\}$ とする（$\sum_i \theta_{ci} = 1$）．なおここで，通常文書分類における変数 d は 1 つを文書を表すが，本研究では 1 つのサイトを表す．また f_i は通常は d 中の単語出現頻度だが，ここではサイト d 中で単語 i を含む記事数とする．

このとき，

$$p(d|\vec{\theta_c}) = \frac{(\sum_i f_i)!}{\prod_i f_i!} \prod_i (\theta_{ci})^{f_i}$$

となる．ここで，θ_c に関する事前分布 $p(\vec{\theta_c})$ を与えると，サイト d がもつ属性値は

$$l(d) = argmax_c \{\log p(\vec{\theta_c}) + \sum_i f_i \log \theta_{ci}\} \qquad (1)$$

と推定できる．

N_{ci} を属性値 c をもつサイトでの単語 i を含む記事の総出現数とすれば，θ_{ci} の最尤推定値 $\hat{\theta}_{ci}$ は

$$\hat{\theta}_{ci} = \frac{N_{ci} + \alpha_i}{\sum_i N_{ci} + \sum_i \alpha_i}$$

と求まる（ただし α_i はスムージングのためのパラメータ）．

3.3 Complement Naive Bayes

Rennie らは [4] で，クラスによって教師データ数に偏りがあることが Multinomial Naive Bayes の精度を下げる主な要因であると指摘されている．その上で，この問題を簡単に回避する方法として，各クラスに「含まれない」確率を学習する方法 Complement Naive Bayes(CNB) を提案している．多くの場合，それぞれの c についてその属性値をもつサイト数のばらつきは，その属性値を「もたない」サイト数のばらつきよりも比率が小さい．このため，補集合を使うことで教師データ数の偏りを抑えられることが多いとされている．

そこで，式 1 の第 2 項の部分を反転させ，CNB では文書 d が属するクラスを

$$l(d) = argmax_c \{\log p(\vec{\theta_c}) - \sum_i f_i \log \theta_{\bar{c}i}\} \qquad (2)$$

と推定する．パラメータ $\theta_{\bar{c}i}$ の最尤推定値は

$$\hat{\theta}_{\bar{c}i} = \frac{N_{\bar{c}i} + \alpha_i}{N_{\bar{c}} + \sum_i \alpha_i} \qquad (3)$$

となる．

3.4 提案手法の概略

まず学習の方法について述べる．教師データはブログサイトのコンテンツとそのサイトの著者の著者属性から成る．これはブログサービスプロバイダの 1 つにお願いして統計情報を使わせてもらったり，ブログ著者向けサービスの登録時にユーザーに入力してもらった値を使うことなどを想定している（本研究では 4. で述べるような方法で表示されている情報をクロールして得た）．まず教師データのあるブログサイトの記事全てを形態素解析器にかけ，bag-of-words モデルでベクトルにする．この際，名詞や動詞のみならず副詞や助詞等を含めた全ての単語を用いる．次に，教師データのあるブログサイトのデータを用い，全ての単語 t について $\chi^2_{max}(t)$ の値を計算し，値が大きい方から一定数の単語集合 W を得る．その後，各属性値 c, 単語 $t \in W$ に対して CNB のパラメータ $\theta_{\bar{c}i}$ を式 3 で推定する（$\alpha_i = 1$ とした）．

次に，著者属性が分からないブログサイトの著者属性を推定する方法について述べる．属性値の推定をする際には，まず推定したいブログサイトの記事を形態素解析器にかけ，単語集合 W に含まれる単語のみから成る bag-of-words ベクトルを作る．その後，式 2 に基づいて属性値を推定する．ここで，式 2 の計算で必要となる $p(\vec{\theta_c})$ について, [4] では一様分布を用いるのがよいとされている．しかし著者属性推定では手がかりがほとんどない場合もあり，このような場合には出力としてその属性の分布そのものが出力されるのが望ましい．このため，本研究では事前分布として一様分布ではなく，教師データ中の各属性値の出現頻度の分布を事前分布をとして用いることとした．

3.5 提案手法の利点

提案手法の既存手法に対する利点は，

- スケーラブルであること
- 汎用的であること

の 2 点である．

まず，CNB 法は計算量が推定・学習共にブログサイト数に対して線形である．素性選択も線形であるためこの手法は全体として線形の計算量，すなわちスケーラブルであるといえる．ラベル付きデータが常時更新され，また教師データ量も必然と多くなるブログにおける属性推定において重要である．後に示す大量の実データを用いた実験でも，問題なく学習・適用が可能であるかが分かる．

もう 1 点は，汎用性である．提案手法では新しい属性を推定させたい場合，サンプルとしていくらかのブログサイトの URL とその属性値を与えるだけで，属性の推定が可能になる．もちろん必ずしも上手く推定可能とは限らないが，テキストから人間が判断できるようなものの多くは提案手法でも推定可能だと考えられる．調査者が属性ごとに適切な手法を探さなくてよいことは，実用上大きな利点であるといえる．

4. 実　　験

本章では，提案手法の有用性を確認するために行った 3 つの実験の手順を説明し，結果を示した上で考察する．提案手法の有用性を確認するための対象として，教師データを入手しやすい性別・居住域・年齢層を対象として用いた．これらの 3 属性のラベル付きデータは「プロフィール欄」などの形でブログに併記されている例があり，またプロフィール欄のフォーマットが決まっている例がいくつかあったため，容易に入手できた．

4.1 実験データ

まず，実験に用いたデータについて紹介する．ブログの URL としては，2006 年 4 月頃から 2007 年 4 月頃までの約 1 年間にクロールした記事の中からブログサイトの URL を抽出したものを用いた．これらのブログのうち，手書きのルールによってプロフィール欄の文字列を取得できたサイトが 13 万 9944 サイトあった（これらのサイトに含まれる記事は 1365 万 4635 件であった）．得られたプロフィール文字列の中には「性別：非公開」といった，意味を成さないものが多く含まれていた．各属性についてこれらの意味の成さないデータを除いたものを学習に使用した．それぞれの属性について意味のある値を得られたサイト数を表 1 に示す．

4.2 素 性 選 択

今回の研究では，著者属性の推定を文書分類問題に帰着させ，機械学習アルゴリズムを適用した．まず，形態素解析(注1)により記事を単語に分解した．次に各属性について，全ての単語 t の $\chi^2_{max}(t)$ を計算した．以降の実験では素性の数を制限しているが，この場合 $\chi^2_{max}(t)$ が大きかった順に素性を選択した．

(注1)：形態素解析には Sen(https://sen.dev.java.net/) を用いた

属性値	サイト数
男性	40,992
女性	42,062
10 代	584
20 代	3,413
30 代	3,036
40 代以降	1,645

属性値	サイト数
北海道/東北	6,654
関東	34,146
中部	9,248
近畿	12,992
中国/四国	4,624
九州/沖縄	5,181

表 1　各属性値を入手できたサイト数

それぞれの属性で $\chi^2_{max}(t)$ 値が大きかったものを表 2 に示す．性別に関しては女性が頻繁に用いるような語が多く並んでおり，分類するにあたり有用な素性を選ぶことができているように見受けられる．居住域については主に地名が並んでいるが，素性の候補を特に地名に限定せずともこのような結果が得られたことは興味深い．また「ほんま」や「やけど」などの方言も含まれており，これは初対面の人であっても方言があれば出身地が分かるという常識と照らし合わせても妥当である．最後の年齢層であるが，学校に関連した語が並んでいる．これは 10 代とその他を識別するのに有用な単語として選択されたものと思われる．10 代を他の年齢層から識別する語は多数候補があるのに対し，他の年齢層（例えば 30 代と 40 代以降など）を識別する語は 10 位以降を見ても少なかった．「妻」や「子供」といった家族の存在を示す語や，「高血圧」など年齢を感じさせる語も存在したが，どちらも決定的な意味を持つものではないといえる．

4.3 精度の測定法

行った実験を説明するのに先立ち，ここで精度の測定法について説明する．今回の研究の出力は各ブログサイトについて各属性値を予測するものであるが，この主な応用先は流行の分析等を想定している．このような目的を考えれば，何 % のサイトについて正しく推定できたかだけでなく，何%の記事について正しく推定できたかも重要である．このため，本研究の実験では既存研究で行われているブログサイト単位での精度の測定に加え，記事単位での精度も測定することとした．

また，推定に当たっては推定をあきらめることはせず，どれかの値を付与した．すなわちすべての実験結果はカバレッジが 100%という条件下での精度である．これは，本研究が流行分析をための著者属性推定という流れで行われているためである．流行分析行う際には属性値を推定しにくいブログでの単語の出現も重要な意味を持つため，属性推定のフェーズで有用な情報を削ってしまうことは望ましくない．このため，推定が困難なサイトについては推定をあきらめるのではなく，流行分析のフェーズで信頼度の低い情報として扱うことを想定した．

性別	
単語	スコア
かしら	6626.84480415172
お天気	5132.94231105452
とっても	5113.70765640604
お花	4937.60301871669
ケーキ	4496.23420212991
ママ	4494.70259723909
美容	4328.24672234456
母	4069.78008474729
俺	3937.26496362870
洋服	3840.90284501630

居住域	
単語	スコア
梅田	7498.625896894276
福岡	6852.738556801302
ほんま	6306.888995094041
札幌	5426.186044104552
新宿	5384.389064902985
大阪	4631.596262013533
熊本	4336.756959634843
渋谷	4104.419198914048
神戸	3960.841161518183
やけど	3924.942990643515

年齢層	
単語	スコア
模	419.6983852720805
教科	405.9396314294655
部活	382.1183746830956
数学	376.5761502449595
体育	311.0252041772567
放課後	308.1818405332941
国語	271.8689324558797
授業	258.8588789909571
テスト	254.4547072312215
自習	250.5657288029923

表 2　各属性で $\chi^2_{max}(t)$ が大きい語

4.4　素性数と精度の関係

第 1 の実験は，各属性について最適な素性数がどの程度かを知ることを目的としたものである．通常の機械学習器の挙動から推測すれば，ある程度までは素性を増やすほど精度が向上するが，それ以降は過学習により徐々に教師データセット以外のデータでの精度が落ちる傾向がある．本研究では，χ^2_{max}(t) 値の大きい方から N 語を文書分類の素性として用いることとしたが，この N を変化させた場合に精度がどのように変化するかを実験的にしらべた．精度は教師データと異なる 4000 サイトのデータで測定した．

結果を表 3 に示す．まず，性別の推定精度については 90%近い精度を得ており，統計データとして用いるに当たっては実用上十分な精度いえる．また居住域の推定についても 8 割以上のブログについて正しく推定できており，実用的な精度を得られているといえる．年齢層については 50%強の精度となっており，あまり高くない．特にサンプルに多い 20 代と 30 代の識別は，人手で書かれたブログから年齢を推定しようとしても，何を手がかりにしてよいかわからない．今回は実験できなかったが，テキストからの書き手の年齢推定は本質的に困難である可能性がある．

全体的に見て，どれも素性数が 1024 のあたりで最も精度が高くなっており，教師データ数とほぼ同数の素性を用いるのがよいことがみてとれる．

4.5　教師データ数と精度の関係

第 2 の実験は，教師データ数と精度の関係をしらべるものである．一般的に学習に用いるデータを増やせば増やすほど精度は向上するが，それに伴い学習に必要となる時間も増えてしまう．新しく入手したデータを用いて頻繁に分類器を再学習するならば，その速度は実用上重要な要素となる．またこの実験は，日々新しいブログデータが来た場合にどの程度の頻度で最学習をすべきかを判断する材料ともなる．この実験では，学習に使用するデータ数を増減させ，精度の変化を調べた．精度は教師データと異なる 4000 サイトのデータで測定した．

結果を表 4 に示す（年齢層の 8192 データでの結果が欠けているのは，年齢層に関して入手できたサンプル数が少なかったためである）．これより，どの属性も教師データが多ければ多いほど精度が高くなる傾向がある．しかし，精度の伸びは鈍化してきている．これより，属性値の種類が数個であれば，教師データは数千件程度あればよいことがわかる．ただ性別と比べて居住域と年齢層では精度が安定するのに必要なデータ量が多くなっており，属性値のとりえる値の数が多ければその分多くの教師データが必要となることを示唆している．

4.6　スケーラビリティの確認

第 3 の実験では，スケーラビリティを確認するため，大量のデータを用いた実験を行った．比較のため，CNB と代表的な高精度文書分類器である SVM で性別推定を行い，学習に要した時間と精度を比較した．これについても精度は教師データと異なる 4000 サイトのデータで測定した．

結果を表 5 に示す．この結果より，大量のデータがある際には CNB は SVM よりも圧倒的に高速になることが分かる．このため，多数の値を取り得る属性の学習をする場合（この場合十分な精度を得るためには多くの教師データが必要になる）や，多数の属性値を学習しなければならない場合は，CNB の利点が発揮されると考えられる．

5. ま と め

本研究では，ブログの著者属性推定問題を単純な文書分類問題と捉え，χ^2 値によって語を選択した上で Complement Naive Bayes を適用する手法を提案した．この手法は，大量の教師データを利用しながらも短時間で学習でき，また個々の著者属性に依存する手法を用いないため任意の属性に適用することができる．

	性別		居住域		年齢層	
素性数	サイト単位	記事単位	サイト単位	記事単位	サイト単位	記事単位
16	65.88%	70.85%	38.60%	52.13%	16.25%	29.38%
32	72.58%	76.09%	47.68%	63.42%	22.45%	35.73%
64	74.55%	78.69%	53.35%	69.02%	36.75%	49.10%
128	80.38%	85.19%	58.43%	72.19%	43.35%	55.36%
256	85.85%	88.10%	62.25%	76.25%	49.30%	58.26%
512	86.78%	88.49%	67.20%	79.63%	49.98%	55.71%
1024	86.80%	88.30%	69.85%	81.48%	50.93%	55.64%
2048	86.28%	87.91%	70.85%	81.15%	50.78%	55.19%
4096	85.85%	87.18%	71.30%	80.06%	49.60%	54.47%
8192	85.15%	86.56%	68.68%	75.61%	50.03%	54.18%

表 3　教師データを一定 (1024 サイト) にして，素性数を変化

	性別		居住域		年齢層	
教師データ数	サイト単位	記事単位	サイト単位	記事単位	サイト単位	記事単位
16	70.30%	69.89%	44.85%	51.09%	36.18%	36.42%
32	76.65%	77.65%	52.63%	60.27%	32.55%	36.71%
64	81.80%	85.00%	54.30%	62.28%	28.20%	30.18%
128	84.33%	85.21%	62.88%	74.40%	40.58%	45.30%
256	84.90%	86.97%	66.65%	78.78%	47.25%	54.47%
512	86.10%	87.79%	68.08%	79.50%	48.38%	54.96%
1024	86.80%	88.30%	69.85%	81.48%	50.93%	55.64%
2048	86.38%	88.71%	69.38%	80.75%	51.38%	57.53%
4096	86.18%	88.01%	70.70%	81.71%	51.58%	56.48%
8192	85.80%	87.71%	71.20%	82.89%		

表 4　素性数を一定 (1024 語) にして，教師データ数を変化

	性別					
	CNB			SVM		
教師データ数	サイト単位	記事単位	時間 (秒)	サイト単位	記事単位	時間 (秒)
1024	86.80%	88.30%	0.10	82.80%	87.46%	1.38
2048	86.38%	88.71%	0.12	81.95%	89.36%	4.60
4096	86.18%	88.01%	0.12	84.33%	91.75%	13.06
8192	85.80%	87.71%	0.19	81.73%	90.94%	38.57
16384	85.98%	87.76%	0.42	81.68%	91.25%	149.15
32768	86.23%	88.02%	0.49	82.10%	91.67%	222.05
65536	86.13%	87.83%	0.88	83.03%	92.02%	419.15

表 5　CNB と SVM の学習速度と精度の変化 (素性は 1024 語)

実験では提案手法を現実のブログデータに適用し，十分な量の教師データ（数クラスの属性であれば数千件程度）があれば文書数と同数程度の素性を用いることで実用的な精度を得られることを確認した．また教師データを大量に用いた場合の速度も高速であることが確認できた．

謝　　辞

本研究で使用したデータの一部は，独立行政法人情報処理推進機構（IPA）「未踏ソフトウェア創造事業」田中二郎 PM による「ブログを用いた『なんでも早期発見システム』の開発」の過程で得たものである．ここに記して感謝します．

文　　献

[1] 安田宜仁，平尾努，鈴木潤，磯崎秀樹：ブログ作者の居住域の推定，言語処理学会第 12 回年次大会 (2006).

[2] 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).

[3] YumingYang，JanO. Pedersen：A Comparative Study on Feature Selection in Text Categorization，Proceedings of the Twentieth International Conference on Machine Learning (1997).

[4] Jason D. M. Rennie，Lawrence Shih，Jaime Teevan，David R. Karger：Tackling the Poor Assumptions of Naive Bayes Text Classifiers，Proceedings of the Twentieth International Conference on Machine Learning (2003).