blogの著者の性別推定

池田 大介† 南野 朋之†† 奥村学‡
†東京工業大学 工学部情報工学科 ††東京工業大学 大学院総合理工学研究科
‡東京工業大学 精密工学研究所
{ikeda, nanno}@lr.pi.titech.ac.jp oku@pi.titech.ac.jp

1 はじめに

近年、blogの急速な普及に伴い、その情報源としての期待が高まってきている。blogは旧来のWebと異なり、リアルタイム性にすぐれ、個人の意見等が多く記述されるという特徴がある。これを利用して、blogからある商品が世間で好評であるか否か、といった評判情報や、現在流行している物は何か、という情報を獲得することを目的とした研究が行われている[2][6]。

こういった研究は主にblog全体を対象に行われており、得られる知識もblog全体での評判や流行である。しかし、現実には、“中年の女性に流行”や、“若い男性には評判がいい”、“小学生に人気であったが大人にも広まる”等、人の属性によって評判や流行に対する傾向が異なることが多い。ここで言う人の属性とは、年齢、性別、職業のような、個人が持つ情報を指す。

こういった傾向の異なりは、マーケティングなどにおいて重要な情報である。例えば、女性向きの商品の広告効果をblogで調査する、といった場合、blog全体でどの程度反響があるか、という情報よりも、広告の対象である女性の評判はどうか、という情報の方が有用であると言える。

既存の研究を利用し、流行や評判の傾向の異なりを捕らえるためには、分析対象となるblogの著者がどういった属性を持っているかが判明していればよい。先の例のように、女性の評判を知りたい場合、どのblogが女性によって書かれたものであるかが判明していれば、それらだけを対象に分析を行うことでこれを知ることができる。

人の属性としては様々なものが考えられる。先述した年齢、性別、職業の他にも、出身地や家族構成なども評判や流行に影響を与えている可能性が高い。

こういった数ある属性のうち、本研究では性別に着目した。性別は、属性の内でも最も評判や流行を左右する要因の一つである。男性向け、女性向けと謳った商品や雑誌が多数存在することもこれを表している。言い換えれば、blogの著者の性別が判明することによって、特徴的で有用な情報が得られると期待できる。

こういった理由から、本研究では、blogからの評判や流行の獲得に利用していくことを目的として、blogの著者の性別の推定を行った。

blogのエントリ中のテキストは口語的に記述されることが多いという特徴がある。そこで本研究では、日本語の話し言葉の特徴を考慮した素性を数種考案した。それらを用い、blogの著者の性別を推定した。

本論文は以下の構成で成る。まず2節で、本研究の関連研究について説明する。次に3節で本論文で提案する性別推定の手法について述べる。続いて4節では提案手法の有効性を確認するための実験と、その際に利用したデータについて述べる。最後に5節で本論文の結論と今後の課題を述べる。

2 関連研究

性別推定 テキストの著者に対する性別推定としては、長文を扱ったもの[3]や、Eメールを扱ったもの[1]等、blog以外のテキストについては研究されている。これらについて紹介する。

前者は、BNCのテキストの著者の性別を推定している。このテキストは平均34000語程度の長文である。素性としては、機能語と品詞、品詞列を使用しており、80%程度の正解率を得ている。また、この結果を利用し、女性は代名詞を好んで使用する、男性は数詞等の数を表す単語を使用する頻度が高い、といった分析を行っている。

後者はSupport Vector Machine(SVM)を用いてEメールの性別の推定を行っている。BNCテキストの性別推定でも使用されているような品詞列や機能語等、言語的な素性に加え、使用されているHTMLタグのようなEメール特有の情報や、空行の数や文の平均長といったテキストの構造に関する素性等、合計200種類以上の素性を利用している。その結果、7割程度の正解率を得ている。

本研究との相違点 これらと本研究との違いは大きく二つ存在する。

一つ目は、先行研究は英語でなされたものであるのに対し、本研究は日本語のテキストを対象としている、という点である。性別の推定には言語的な特徴が大変有効である。そこで本研究では、日本語における男女差を考慮した素性を提案する。

もう一つは、本研究の対象がblogであるという点である。次節で詳しく述べるが、blogには著者の性別推定をするにあたり、BNCテキストやEメールとは異なる性質がいくつか存在する。これによって、blogの著者の性別推定は他の種類のテキストに対するそれとは問題の質が若干異なる。

3 性別推定の手法

3.1 概要

本研究では、blogの著者を“男性”、“女性”、“性別不明”の3つのクラスに分類する。確信度を用い、それが十分に大きい事例だけを“男性”、“女性”のクラスに分類し、そうで無い事例は“性別不明”クラスとする。分類器としてはSVMを用い、確信度にはSVMの出力する分離超平面からの距離を用いた。

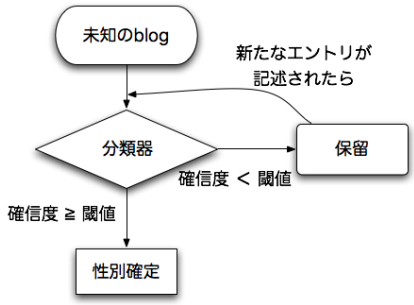

図 1: システムの概要

“性別不明” クラスを用意する理由は大きく以下の 3 つである。これらは blog 特有の性質であり、これによって他の種類のテキストの著者の性別推定とは問題の質が若干異なってくる。

人手でも判断が難しい blog、正解の存在しない blog が存在する　blog によっては、人間が見ても著者の性別が判断できない場合がある。このような事例を機械に判断させるのは難しい。また、企業やサークルのような団体で運営される blog もあり、こういった blog はそもそも性別の推定自体が無意味である。

再現率よりも精度を重視したい　現在、日本には何十万もの blog が存在する。この数は評判情報や流行の獲得には十分である。再現率が多少低くとも、精度が確保できていれば、そこから得られる知識も精度の高い物となると考えられる。

現存するエントリだけで判断する必要は無い　blog はその著者によって日々更新される。そのため、現存するエントリのみから性別が判断できなくとも、その後のエントリを見る事で判断できるようになる可能性がある。つまり、その時点では判断が難しいような blog は “性別不明” クラスに分類しておき、後にエントリが追加されてからもう一度判断する。

図 1 にシステムの流れを示す。

本節の残りでは分類の際に使用した素性について説明する。

3.2　素性

本研究では blog エントリ中のテキストから得られる素性のみを使用した。blog にはエントリ以外にも画像を始め様々な情報が存在する。しかし、これらの解析には、blog の複雑な構造も手伝い、多大なコストがかかる。

我々は、“テキストの情報だけでも、性別は十分に推定できる” と予想し、テキストから得られる素性のみを使用した。また、この予想の検証は 4 節で行う。

また、“甘いお菓子は女性の方が好む” など、男女では大きく趣向が異なり、こういった意味論に踏み入った素性は有用である可能性は高い。しかし、本研究では、性別を推定した結果を blog からの評判や流行の情報の獲得に適用することを考えている。そのため、性別を推定する段階でそれらの情報を素性として利用すると、そこから得られる結果もその素性に依存したものになってしまうと考えられる。そのため、本研究ではこういった意味論に踏み込んだ素性は用いなかった。

一人称代名詞　先述した通り、blog では日記のように書き手自身の経験や意見を述べられることが多い。つまり、blog においての一人称は書き手自身であることが多く、その場合の一人称代名詞は著者が自分を指す際に使用しているものであると考えられる。

日本語、特に話し言葉では男女で使用される一人称代名詞の傾向は大きく異なる。直感的にも、“俺”、“僕” 等は主に男性が使用し、“あたし” 等は女性の方が使用頻度が高い。そのため、blog で主に使用されている一人称代名詞は性別の推定に有効であると考えられる。

blog 中には、主に使用している一人称以外にも引用文や会話文などで様々な一人称が現れる。こういったものは、むしろノイズになってしまうと考えられる。そこで、本研究では blog 中に最も多く現れる一人称代名詞が、その blog の著者が自分を指す際に使用する代名詞であると考え、これを素性とした。候補となる一人称代名詞は IPADIC の “代名詞-一般” の項より人手で選択した。

機能語　関連研究でも挙げた通り、著者推定 [4] や性別推定といったタスクでは機能語の出現頻度が素性として有効な事が多い。こういった研究は主に英語で行われており、日本語の著者推定では機能語のみではなく形態素全てを素性として使用した方がよいという報告もある [9]。しかし、語尾の違い等を考えると、性別の推定というタスクにおいては、形態素全てより機能語を利用した方が良い結果が得られると考えられる。

男女の話し言葉には、特に終助詞において差があると指摘されている [8]。この研究では、大学生の男女の会話を録音し、それを分析している。blog は完全な話し言葉ではないが、先に述べた通り話し言葉的な書かれ方をする事が多い。そこで、本研究でも関連研究同様、各機能語の 1 エントリあたりの出現頻度を素性として組み込んだ。

形態素　機能語を素性として利用することを述べたが、自立語にも性別の推定に有効と思われるものは少なくない。一部の語は男性と女性で使用頻度が大きく異なる。例えば、“飯 (めし)” という語は男性的な印象を受ける。逆に “かわいい” 等は女性の方が好む。つまり、“かわいい” という語を使用している blog の著者は女性の可能性が高く、推定に有用であると考えられる。

こういった語を選定するために、本研究では χ 二乗値を用いた。χ 二乗値は、有意検定をする際によく使用される値であり、素性選択の方法として有効であると報告されている [5]。blog 中に現れる全形態素に対し、男女のクラスで χ 二乗値を求め、その大きい物から一定数を素性とした。実験に使用したデータを用いて求めた χ 二乗値の特に大きい形態素の一部とその χ 二乗値を表 1 に示す。

4　実験

本研究では、提案手法の有効性を確認するため、いくつかの実験を行った。本節ではそれらについて報告する。SVM の実装として TinySVM を使用し、5 分割交差検定による評価を行った。カーネルとしては線形

表 1: χ 二乗値の大きい形態素の例

χ 二乗値	形態素
89.6188	私
50.6925	ちゃん
42.5347	かしら
40.0182	買い物
39.8401	もらう
38.4655	友達
35.8619	ちゃんと
34.3010	とっても

カーネルを用い、コストマージンパラメータ C は、それぞれ結果が最良になるものを求め、その結果で評価した。

4.1 データについて

実験の際に使用したデータについて説明する。

データの性質 一部の blog ホスティングサービスでは、自分の年齢や性別といったプロフィールを記述する欄が用意されている。本研究ではここに性別の記述がある blog をデータとして用いた。そのため、全ての blog に “男性”、“女性” のどちらかのラベルが付与されている。プロフィール欄の記述を利用する利点として、人手でラベル付けを行うのに比べコストが低く、容易に学習のためのデータを増やせる、という点が挙げられる。

本研究では、Yahoo!blog、楽天 blog から収集した上記の条件に沿う blog 計 612 件を使用した。各 blog からは 10 件ずつエントリを収集し、それらを推定に利用した。データのうち、“男性” ラベルの付いたものが 264 件、“女性” ラベルの付いた物が 348 件あった。

人手によるラベル付け このデータに対し、人手によるラベル付けを行った。ラベル付けの際には、blog 中の画像やサイドバーといった情報を利用せず、エントリのテキストのみから性別を推定した。ラベル付けは一人の主観で行った。本論文で提案するシステム同様、推定できなかった物は “性別不明” とした。

すでにラベルの付いているデータにもう一度ラベル付けを行う目的は、プロフィール欄の記述が人間の付けるラベルに十分近い事を確認することと、画像等の情報を用いずとも、テキストだけで性別の推定が可能である、という仮説の検証をすることである。

人手で付けたラベルと、プロフィール欄の記述との関係を表 2 に示す。表中の male、female はそれぞれ “男性”、“女性” のクラス、unknown は “性別不明” クラスに分類されたことを意味する。

“性別不明” を除くと、人手によるラベルとプロフィール欄の記述は 95.3%一致しており、プロフィール欄の記述は十分に人手のラベルに近いと言える。また、全データの 86.6%の blog に対し、性別が推定できており、テキストの情報だけでも性別の推定が十分可能であった。

前処理 収集したエントリから、人手で作成したルールを用いて本文部分のテキストのみを抽出した。それらを文区切りの後、形態素解析を行った。文区切りには須山らの手法 [7] を用い、形態素解析には ChaSen を利用した。

表 2: 人手の結果とプロフィール欄の記述との比較

		プロフィール欄	
		male	female
人手	male	190	13
	female	12	315
	unknown	62	20

表 3: 素性の組み合わせの実験

素性セット	accuracy
機能語 + 一人称	0.833
機能語 + 一人称 + 形態素 10	0.875
機能語 + 一人称 + 形態素 50	**0.889**
機能語 + 一人称 + 形態素 100	0.874
機能語 + 一人称 + 形態素 500	0.855
機能語 + 一人称 + 全形態素	0.856
一人称 + 形態素 50	0.877
機能語 + 一人称 + 形態素 50	**0.889**
形態素頻度 + 一人称 + 形態素 50	0.874
機能語 + 一人称 + 形態素 50	**0.889**
機能語 + 形態素 50	0.859

4.2 男女 2 クラスでの実験

まず、本論文で提案した素性の有効性を確認するため、“男性”、“女性” の 2 クラスに、SVM の出力通り分類する実験を行った。評価尺度には全データのうちの正解したデータの割合を示す正解率 (accuracy) を用いた。

提案手法では形態素を素性として利用する際、χ 二乗値を用いた素性選択を行う。このとき選択する素性数を変化させ、その最適な数を求めた。形態素以外の素性としては、一人称代名詞と機能語の頻度を用いた。

表 3 上段がその結果である。形態素の素性として χ 二乗値の上位 50 語を利用した素性セットが最も良い結果を得た。逆に、形態素を素性として使用しないものが最も悪い結果となり、次いで素性選択を行わない物の結果が悪かった。このことから、形態素を χ 二乗値を用いて絞り込んだ素性が有効であったと言える。

次に、機能語の頻度、一人称代名詞の素性としての有効性を確認するため、それぞれを素性セットから削除したものと比較した。機能語の頻度に対しては、提案手法である機能語のみを用いた素性と、全形態素の頻度を素性とした場合とを比較した。

それぞれの結果は表 3 の中段、下段に示す。どの組み合わせの素性セットも、提案手法のそれを下回る結果となった。この結果から、本論文で提案した素性が性別推定に有効であったと言える。

4.3 “性別不明” クラスを含めた実験

次に、SVM の出力である分離平面からの距離に閾値を設け、それを下回るものは “性別不明” クラスに分類する実験を行った。

まず、閾値を変えて “男性” クラス、“女性” クラス

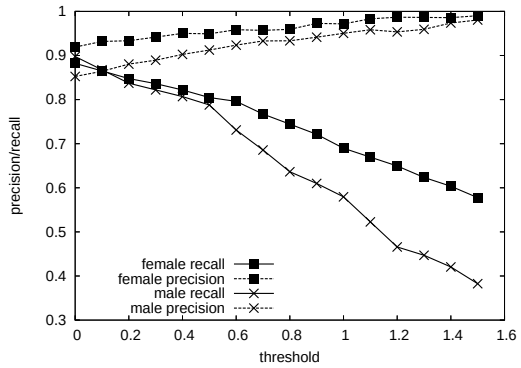

図 2: 閾値を変化させたときの精度と再現率

の再現率 (recall)、精度 (precision) がどう変化するかを調べた。素性としては、前実験で最も良い結果を得た、一人称代名詞、機能語、形態素 50 語のセットを使用した。図 2 は、閾値を変化させたときのそれぞれのクラスの精度、再現率の推移を示した物である。

この図から、“男性” クラスの推定に比べ、“女性” クラスの推定の方が精度、再現率とも高い値が得られる事がわかる。これは、本研究で用いた素性が、いずれも男女のテキストのスタイルの差を捉えたものであることに起因する。男性は女性に比べ、書き言葉的な文書や特徴の無いスタイルで blog を記述することが多く、本研究で用いた素性ではそういったテキストからは性別の推定は難しい。

閾値を変化させた結果、閾値 0.5 のとき、“男性” クラスでは再現率 0.79 に対し精度 0.91 を、“女性” クラスでは再現率 0.81 に対し精度 0.95 と、再現率を大きく損なう事無く高い精度を得る事ができた。この精度ならば、blog からの評判や流行の分析にも十分役立てる事が可能である。

次に、システムの出力したクラスと、人手でつけたラベルとの比較を行った。表 4 がその結果である。システムは閾値 0.5 を使用した際の結果で、表中の括弧内の結果は、閾値を使用せず、“男性”、“女性” の 2 クラスに分類した場合の結果である。

理想的な結果は、システムが “性別不明” クラスに分類する事例は、人手でも “性別不明” であるか、人手とシステムで判断が食い違うものだけ、という結果である。今回の結果でも、人手とシステムで判断が食い違う物は、3 クラスにすることによって、大部分が “性別不明” クラスに分類された。しかし、システムが “性別不明” としたもので、人手でも “性別不明” と判断されている事例は 15 件しかなく、ほとんど一致しなかった。

この理由は、人間とシステムの “性別不明” に、性質の違いがあるためである。人手における “性別不明” には、団体によって運営されている blog や、リンク集のような blog など、blog の著者の個人的な事の書かれない blog が数多く含まれている。こういった物でも、テキストに特徴があればシステムは性別を推定してしまう。逆に、特徴の無いテキストはシステムは “性別不明” に分類しやすい。テキスト中に、人間なら性別を確定できるような記述が存在しても、システムはその情報を使用する事ができない。

表 4: 人手の結果とシステムの出力との比較

		システム		
		male	female	unknown
人手	male	152(175)	18(28)	33(0)
	female	17(34)	269(293)	41(0)
	unknown	59(69)	8(13)	15(0)

5 結論

本研究では性別の推定を行い、blog を “男性”、“女性”、“性別不明” の 3 つのクラスに分類した。日本語における男女の話し言葉の性質の違いを考慮した素性を利用する事で、高い精度を得た。

しかし、人手でラベル付けした結果と比較すると、“性別不明” クラスではまだ差が大きい。今後はこの差を埋め、より人手によるラベルに近い結果を得られるシステムを目指したい。具体的には、団体で運営されている blog 等、性別の推定をする意味の無い blog を自動的に判断し、あらかじめ “性別不明” クラスと決めてしまう、という手法を考えている。

また、職業や年齢と行った、性別以外の属性の推定にも挑戦していきたい。

参考文献

[1] Malcolm Corney, Olivier de Vel, Alison Anderson, and Georoge Mohay. Gender-preferential text mining of e-mail discourse. In *18th Annual Computer Security Applications Conference*, 2002.

[2] Toshiaki FUJIKI, Tomoyuki NANNO, Yasuhiro SUZUKI, and Manabu OKUMURA. Identification of bursts in a document stream. In *First International Workshop on Knowledge Discovery in Data Streams*, 2004.

[3] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Liguistic Computing*, Vol. 17, No. 4, 2003.

[4] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, Vol. 26, No. 4, pp. 471–495, 2000.

[5] Yuming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, pp. 412–420, 1997.

[6] 鈴木泰裕, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. 人工知能学会, セマンティックウェブとオントロジー研究会, 2004.

[7] 須山敦. 機械学習による html 文書の文特定. 東京工業大学大学院総合理工学研究科知能システム科学専攻修士論文, 2005.

[8] 小早川百合. 話し言葉の男女差 -定義・意識・実際-. 日本語ジェンダー学会誌 4 号, 2004.

[9] 坪井祐太, 松本裕治. Authorship identification for heterogenuous documents. 情報処理学会研究報告 自然言語処理研究会, Vol. 2002-NL-148, pp. 17–24, 2002.