3D2-04

# ブログ分類知識に基づく男性語･女性語の抽出

## Extraction of Gender Word Based on Classification Knowledge of Weblog

小林大祐[*1] 松村真宏[*2] 石塚満[*1]
Daisuke Kobayashi Naohiro Matsumura Mitsuru Ishizuka

[*1] 東京大学大学院情報理工学系研究科 [*2] 大阪大学大学院経済学研究科
The University of Tokyo Osaka University

We use and correspond with the document on Doblog and a result of questionnaire survey for Doblog user, and we structure a system for classifying blog document into either male or female. And we pick out the keyword from each of non-division document, document written by male or female, and document classified by this system into male or female, and investigate distinction of each keyword. As a result, the system can be used for web advertisement, and for getting estimation information with distinction of gender.

## 1. はじめに

日々更新されるインターネット上のブログには、ユーザ個人の意見がすばやく反映され、個人の意見が多く記述されるという特徴がある。

ブログ記事から現在のトレンドを汲み取り、それを新商品開発や宣伝に活用する流れは、現在のマーケティング活動の中で大きな影響を及ぼしつつある。

しかし、ブログ記事の書き手の個人情報は一般的には公開されていない。ブログ記事の書き手が男性であるか女性であるかを判別できるようになれば、マーケティング活動でも特に有用なものとなる。

また、計算機によって、文章から N-Gram を抽出し、その表現の差を見ることで、書き手の性別により異なる単語を使用するという発見があった[近藤 01]。

本研究では、Doblog[1] におけるブログ記事とそのユーザに対して行った大規模なアンケート調査結果を対応づけて利用し、ブログ記事の文章から書き手の性別を判別するシステムを構築した。

以下、2 章ではブログ記事の書き手を男女分類する本研究のシステムについて、3 章では男性がよく用いる語・女性がよく用いる語の抽出について述べ、4 章で実験を行い、5 章でまとめる。

## 2. ブログ記事の性別分類

### 2.1 基本的な手法

まず、各ブログ記事の素性ベクトルを生成する。素性として、各記事を形態素解析し、N-Gram を抽出する。形態素解析器としては chasen を用いた。N-Gram を構成する品詞は、名詞、形容詞、動詞とする。N は 10 以下とした。

$$tfidf(t,D) = tf(t,D) \times idf(t)$$

$tf(t,D)$ = ブログ記事Dに単語$t$が出てくる回数

$$idf(t) = \log\left(\frac{N}{df(t)} + 1\right)$$

$df(t)$ = 単語$t$が出てくるブログ記事の数

$N$ = 全ブログ記事数

連絡先：小林大祐, 東京大学情報理工学系研究科, 東京都文京区本郷 7-3-1, d-koba@mi.ci.i.u-tokyo.ac.jp

各素性の特徴量としては、tfidf を用いる。今回用いる素性は、tf の合計が 3 以上のものとする。df については制限を設けない。

これを正規化することにより、各ブログ記事の素性ベクトルが生成される。分類器としては、非線形 SVM を用いる。

### 2.2 分類困難な文章のフィルタリング

しかし、単にこのまま分類しただけでは、文章の量の短い記事も存在するなど、精度があまりよくないことが想像される。

そこで、より分類を確実にするために、男女分類が困難であると思われる文章をフィルタリングする。これにより、フィルタリングされないブログ記事は、より確実に男性の書いた文章であるか女性の書いた文章であるか分類することができるようになる。

#### (1) 分類困難な文章

どのような文章が分類困難であるかということを考える。一つは、ブログ記事が分類境界面に近いと考えられる記事である。もう一つは、ブログ記事に含まれる単語の種類が極端に少なく、分類しづらいと考えられる記事である。

これらの 2 種類の文章を事前にフィルタリングするために、各ブログ記事から男性度･女性度というものを求める。さらに、男性度･女性度をもとめるためには、各素性の重みが必要となるので、それについて説明を行う。

#### (2) 各素性の重み抽出

線形 SVM を用いてのキーワード抽出を行う[Brank 02]。この手法は、始めに線形 SVM 分類器を学習させ、単語の重みを抽出する。こうして抽出された重みは、それぞれの単語がどの程度分類に効いているか、また、正例、負例のどちらかに効いているかを示している。

線形 SVM を用いるのは、重みが簡単に計算できるためである。

$$prediction(\mathbf{x}) = \mathrm{sgn}[b + \mathbf{w}^{\mathrm{T}}\mathbf{x}]$$

以上が線形 SVM を表す式である。$\mathbf{x}$は入力された素性ベクトル、$\mathbf{w}$は重みベクトル、b はパラメータである。素性ベクトルとして以下のものを入力すれば、素性 t に対応する重みを求めることができる。なお、b=0 となるように学習させると、計算を簡単にすることができる。

[1] http://www.doblog.com/

[2] なお、本研究では、男女に関する固定的イメージや性別による固定的な役割分担意識の存在を仮定しているのでない。

$$\mathbf{x} = (x_1, x_2, \ldots, x_d)$$

$$x_i = \begin{cases} 1 & \text{if } i = t \\ 0 & \text{if } i \neq t \end{cases}$$

このようにして求めた素性 t の重みをw(t)と置く。このw(t)が正である場合は正例を表す素性として、負である場合は負である素性として抽出される。

#### (3) 各ブログ記事の男性度･女性度計算

これを用いて各ブログ記事の男性度･女性度を計算する。なお、この論文では、女性が書いたブログ記事を正例、男性が書いたブログ記事を負例としている。

ここでは、ブログ記事 b の男性度･女性度というものを、以下のように定義する。

$W_b$ ＝ブログ記事**b**に含まれる素性の集合

$$s_f^b = \sum_{t \in T_f} w(t)$$

$$T_f = \{\forall t \mid t \in W_b, w(t) > 0\}$$

$$s_m^b = -\sum_{t \in T_m} w(t)$$

$$T_m = \{\forall t \mid t \in W_b, w(t) < 0\}$$

ここで求めた、$s_f^b$をブログ記事 b の女性度、$s_m^b$をブログ記事 b の男性度とする。

#### (4) 男性度･女性度によるフィルタリング

フィルタリング方法として 2 種類の方法が考えられる。1 つはブログ記事に含まれている単語の数が少ない場合である。単語の数が少ない場合は、男女分類が困難になると思われるためである。これを解決するため、以下の閾値 $c_s$ を導入する。男性度･女性度がともにこれを下回っている場合は、男女分類が困難だと判断する。

$$s_m^b \geq c_s,\ s_f^b \geq c_s$$

もう1つは、男性度と女性度の差が少ない場合である。これについては以下の閾値 $c_g$ を導入する。閾値を下回っている場合に男女分類が困難だと判定する。

$$|\log(s_f^b / s_m^b)| \geq c_g$$

### 2.3 男女分類システムのオーバーオール

分類を行う前に各単語の重みをあらかじめ算出する。入力された文章はまず形態素解析され、文章に含まれている単語と重み、それに応じた男性度･女性度が計算される。この男性度･女性度を用いて分類困難な文章をフィルタリングする。

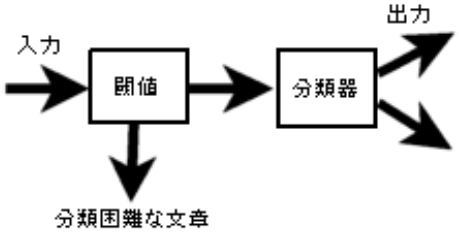


図 1　本システムの構成

表 1　本稿「1.はじめに」中の男性語・女性語

| 女性語 | 重み | 男性語 | 重み |
|---|---|---|---|
| 章 | **0.708** | それ | **-1.241** |
| 1 | **0.485** | する | **-1.174** |
| 中 | **0.439** | 影響 | **-0.793** |
| 表現 | **0.394** | もの | **-0.716** |
| 宣伝 | **0.298** | する れる | **-0.665** |
| 文章 | **0.294** | 記事 | **-0.556** |
| 行く | **0.281** | ユーザ | **-0.474** |
| 男性 | **0.265** | 女性 | **-0.390** |
| 差 | **0.148** | 研究 | **-0.363** |
| 規模 | **0.104** | 現在 | **-0.317** |
| 女性度 | **3.613** | 男性度 | **9.144** |

そして、フィルタリングされずに残った文章を、SVM の分類器にかけ、ブログ記事の書き手の男女分類を行う。図 1 に簡単な図を載せてある。

このため、単に分類器の精度だけではなく、どれだけフィルタリングされる文書の数を抑えたまま精度を上げるか、ということが非常に重要になる。

## 3. 男性語･女性語の抽出

2.2.1 節で求めた各素性の重みは、そのまま各素性となる単語を、どの程度男性または女性が使用しているかという指標にもなると考えられる。ここでは、重みが正である単語を女性語、重みが負である単語を男性語として扱う。

これが妥当であるか評価するために、男女を区別しないブログ記事、男性が書いたとラベル付けしてあるブログ記事、女性が書いたとラベル付けしてあるブログ記事、本研究のシステムで男性だと分類されたブログ記事、本研究のシステムで女性だと推測されたブログ記事、5 種類それぞれからキーワードを取り出し、得られるキーワードの違いについて検討する。

### 3.1 文章からの男性語･女性語の抽出

男性語･女性語を実際の文章から抜き出す例として、本論文の「1.はじめに」から男性語･女性語を抽出し、男性度･女性度を求めてみることとした。

この結果は表 1 に載せた。これによると、この文章は男性度(9.14)が女性度(3.16)に比べて高いということが分かる。

## 4. 実験

今回実験に用いるブログ記事は、Doblog がユーザに対して行った大規模なアンケート調査結果により書き手の性別が判別された記事 241,251 件(男性の記事 157,817 件、女性の記事 83,434 件)のうち、学習データとして男女それぞれ 1,000 件(合計 2,000 件)、テストデータとして学習データとは別に男女それぞれ 10,000 件(合計 20,000 件)のブログ記事を用いた。そして、それぞれのブログ記事からタグを除去した。

### 4.1 各素性の重みの抽出

まずは素性のリストを作成する。このリストは学習データ、テストデータをして取得した全ての文章から、tf が 3 以上の N-Gram を抽出する。この時、素性の数は 67,237 となった。

この各素性に対して、学習データより線形 SVM 学習器を作り、重みを抽出した。この絶対値が大きかった素性を表 2 に載せる。この重みが大きい単語は、一人称を表す素性が多かった。

表 2　各素性の重み

| 素性 | 重み | 素性 | 重み |
|---|---|---|---|
| 私 | **3.93** | 僕 | **-2.78** |
| 笑 | **2.31** | 俺 | **-2.51** |
| あたし | **2.13** | ・ | **-1.84** |
| わたし | **1.86** | 0 | **-1.59** |
| 女 | **1.79** | いうこと | **-1.37** |
| 名前 | **1.78** | 東京 | **-1.35** |
| 食べる | **1.41** | 曲 | **-1.34** |
| コト | **1.40** | こちら | **-1.33** |
| 今度 | **1.38** | それ | **-1.24** |
| ワタシ | **1.37** | 書く | **-1.19** |

## 4.2 フィルタリングの閾値

### (1) 2 つの閾値の変化による精度･フィルタリングされるブログ記事数の変化

この素性の重みを利用して、男女分類が困難であるブログ記事をフィルタリングする。なお、まったくフィルタリングがない場合の分類器の F 値は 0.663 であった。これに対し、2.2.3 節で述べた 2 つの閾値を利用して男女分類の精度が上がるかどうか、また、フィルタリングされる文書数が増えすぎないかの実験を行う。まず、男性度･女性度による閾値 $c_s$ のみを変化させた時の Precision、Recall、F 値、Accuracy の変化が図 2 である。これを見ると、閾値を上げるに伴って、残り記事の分類の精度は上がるが、分類できるブログ記事が減少することが分かる。

また、同様の変化を、今度は男性度と女性度の比の閾値 $c_g$ のみを変化させて精度･分類できるブログ記事数がどうなるかについて見た。図 3 によると閾値を変化させた時に分類できる記事数が減少するのは図 2 と同様だが、図 2 とは違って、閾値を上げすぎると精度が下がってしまうことが分かる。

次に、2 つの閾値を同時に設定した場合にどうなるかを見る。この 2 つの閾値を変化させた際に、分類できる記事の数と F 値がどうなるかを調べたのが図 4 である。これによると、残り記事数が減るにつれて適切な閾値を設定してやることにより、より高い精度で分類ができるようになることがわかる。

### (2) フィルタリングの閾値の決定

実際に F=0.8 などとなる 2 つの閾値の設定の中で、フィルタリングされないで残る文章数の最大値と、それを与える 2 つの閾値がいくらかを実際に調べることとする。

男性度と女性度の比による閾値 $c_g$ を少しずつ変化させていき、それぞれの値において、もう 1 つの男性度･女性度の閾値 $c_s$ を変化させていって、ある F 値になるときの男性度･女性度の閾値 $c_s$ と、分類できるブログ記事の数を調べる。この時の様子が図 5、図 6 である。

図 5 は、男性度と女性度の比の閾値 $c_g$ と、その時にある F 値を始めて超える男性度･女性度の閾値 $c_s$ の関係を表している。これによると、2 つの閾値をあげれば、共に分類の精度が上がることが分かる。

また、片方の閾値を下げた場合は、もう片方の閾値を上げないと精度が上がらない、という点も分かる。これにより、2 つの閾値がトレードオフになっていることが分かる。

図 6 は、男性度と女性度の比の閾値 $c_g$ と、分類できるブログ記事の数の関係を表している。$c_g$ を上げた場合は、フィルタリングされる記事の数が多くなるので、残り記事数も少なくなる。また、$c_g$ が小さい場合でも、図 5 で見たように 2 つの閾値がトレードオフの関係となっているので $c_s$ が大きくなり、結果的にフィルタリングされる記事の数が多くなってしまう。

よって、2 つの閾値をそこそこに設定した場合に、フィルタリングされないで残るブログ記事の数が最も多い、という結果が得られる。

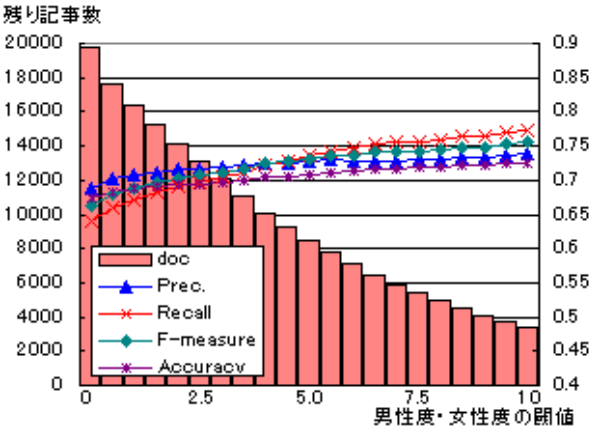


図 2 $c_s$ による、精度と
フィルタリングされずに残るブログ記事数の変化

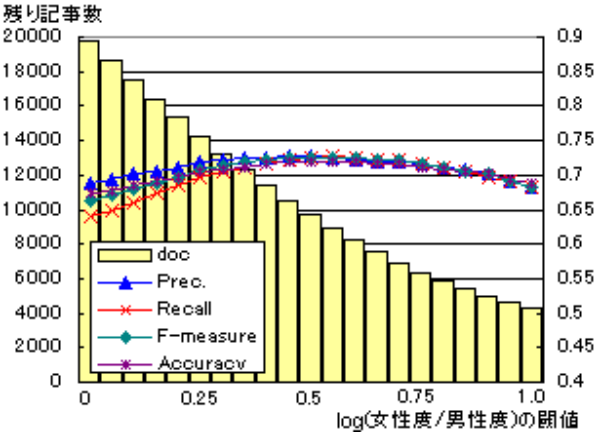


図 3 $c_g$ による、精度と
フィルタリングされずに残るブログ記事数の変化

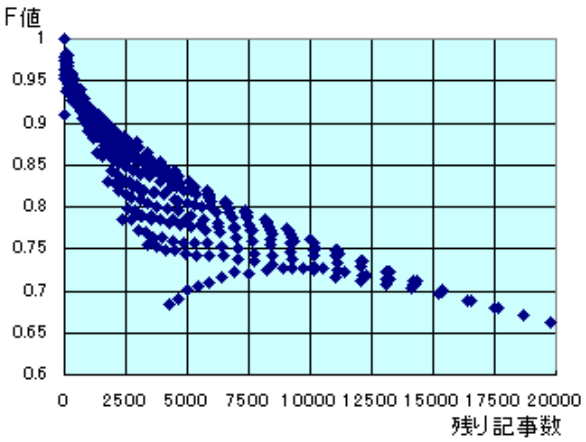


図 4　2 つの閾値を変化させた時の F 値と
フィルタリングされずに残るブログ記事数の関係

表 3　各キーワードの男性,女性,混合文章による tfidf 値

| Keyword | Female | Mixed | Male |
|---|---|---|---|
| 彼氏 | 255 | 313 | 6 |
| 彼女 | 143 | 374 | 211 |
| 夫婦 | 292 | 396 | 280 |
| 両親 | 10 | 85 | 6 |
| 母親 | 15 | 50 | 67 |
| 父親 | 15 | 27 | 9 |
| 親子 | 77 | 160 | 26 |
| 親戚 | 117 | 199 | 2 |
| 親族 | 0 | 91 | 122 |
| 親友 | 78 | 77 | 1 |
| 友達 | 208 | 232 | 165 |
| 友達 関係 | 1065 | 1881 | 0 |
| いい 友達 | 107 | 107 | 0 |
| 食べる | 795 | 1662 | 1084 |
| 全部 食べる | 1713 | 4027 | 0 |
| 食べる　られる | 62 | 284 | 205 |
| 歩く | 122 | 235 | 184 |
| 走る | 71 | 0 | 0 |
| 考える | 191 | 398 | 244 |
| 出かける | 232 | 334 | 76 |
| 買う | 530 | 1321 | 942 |
| 売る | 96 | 187 | 160 |

### 4.3 男性語･女性語の抽出

男性の文章から抽出されたキーワード、女性の文章から抽出されたキーワードを比較することにより、男性がよく使う語･女性がよく使う語を抽出することができる。様々な種類の単語において、これを抽出する事を試みる。これらのの tfidf 値の一覧は表 3 に載せてある。

まず、友人関係を表す単語は「友達」「親友」など、総じて女性文章から抽出された tfidf 値が高かった。それ以外の人間関係ついて表す単語は、「夫婦」「両親」「父親」などは男性文章･女性文章の tfidf 値はあまり変わらないが、「彼氏」「親子」「親戚」は女性文章の tfidf 値が高く、逆に「彼女」「母親」「親族」は男性文章の tfidf 値が高いことが分かる。

例えば「彼氏」「彼女」は付き合っている人を指す言葉であり、男女どちらか一方の文章からよく抽出されるのは容易に想像できることである。しかし、「親戚」「親族」はほとんど同一の内容をさす言葉であるが、それぞれ男性･女性の片方の文章からよく抽出された。これは、同じものをさす場合でも、男女で使用する単語が異なることがある、ということを示している。

最後に動詞について調べてみた。男性も女性もとることができる行動は同じため、大部分の動詞で男女の tfidf 値がほとんど同じになると予想される。しかし、実際には「走る」「出かける」などといった動詞は女性からよく抽出され、「買う」「売る」といった動詞は男性からよく抽出された。

さらに、「食べる」などの単語はあまり男女の差は見られなかったが、「全部食べる」や「食べられる」など語尾が活用したり、名詞などの単語がくっついたりした場合に、男女どちらかで現れる、という傾向が顕著になる場合も見られた。

## 5. まとめ

本稿では、ブログ記事を高い精度で分類できることが示された。また、その過程で男性がよく使う単語、女性がよく使う単語を抽出することができた。

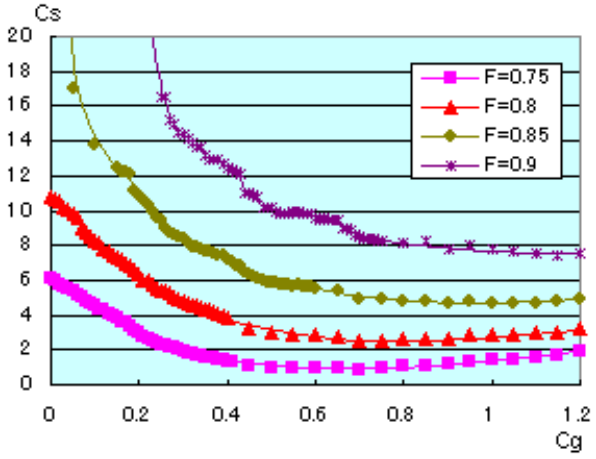


図 5　2 つの閾値と F 値

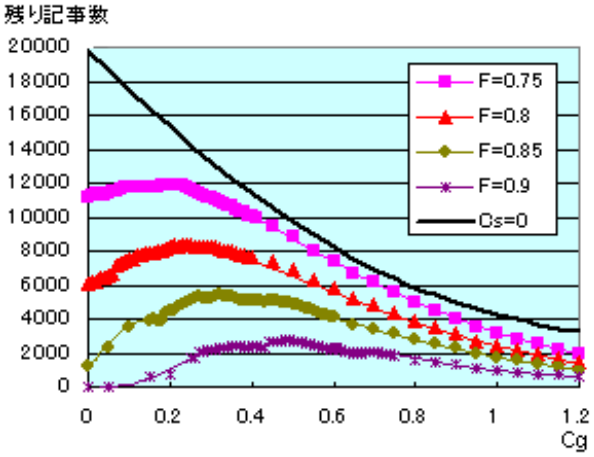


図 6　一定の F 値による $C_g$ と
フィルタリングされずに残る文章数の関係

4.3 節の単語の抽出に関しても、今回用いた素性が、自立語の名詞、動詞、形容詞のみであった。そのため、男女で動詞のあとに助動詞や助詞などがつくことで構成される文末の表現が違うということが考え付くが、今回はその表現の違いについては抽出できなかった。今後さらに詳しく研究したい。

今後は更に精度を高く、またより多くのブログ記事を分類することや、本手法の応用として年代別のブログ記事の分類、年代ごとによく使う単語の抽出などを試みる。

## 謝辞

本研究では Doblog のブログ記事のデータおよび、「Doblog の利用に関するアンケート調査」の結果のデータ提供をいただき、分析に利用させていただきました。協力していただきました株式会社 NTT データ様、株式会社ホットリンク様には記してお礼を申し上げます。

## 参考文献

[近藤 01]　近藤 泰弘,近藤 みゆき．「平安時代古典語古典文学研究のための N-gram を用いた解析手法」（言語情報処理学会第 7 回年次大会『発表論文集』2001）

[Brank 02]　Janez Brank, Marko Grobelnik, Nataša Milić-Frayling, and Dunja Mladenić. Feature selection using support vector machines. Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields, Bologna, Italy, September 2002.