2F4-09

ブログ記事における男女別・年代別・地域別傾向の分析

Analysis of Gender-, Age-, and Area-Specific Trend in Blog Articles

松村 真宏[*1]　三浦 麻子[*2]
Naohiro Matsumura　Asako Miura

[*1]大阪大学大学院経済学研究科
Graduate School of Economics, Osaka University

[*2]神戸学院大学人文学部
Faculty of Humanities and Sciences, Kobe Gakuin University

In this paper, we explore blog articles to understand gender-, age-, and area-specific trend as a part of blog modernology project. We first classify blog articles into gender-, age-, and area-related classes using Naive Bayes Classifier with two thresholding approaches, and achieve more than 20% higher accuracy than baseline approach. Then, from the learnt models from Naive Bayes Classifier, we extract characteristic features and understand our lifestyle from gender, age, and area point of view.

1. はじめに

ブログ記事は人々の日常が綴られたテキストアーカイブであり，そこから人々の行動や思考の軌跡を垣間見ることができる．筆者らは，ブログ記事を通して都市風俗を観察するブログ考現学プロジェクトを進めており [松村 06]，その一環として本稿では，機械学習の手法を用いて男女別・地域別・年代別傾向の考現学的な分析を試みる．

2. 関連研究

野呂らはブログ記事中に書かれているイベントの抽出および生起時間帯をナイーブベイズと EM アルゴリズムを組み合わせて抽出している [野呂 06]．H. Liu らはナイーブベイズを用いて男女分類に特徴的な語を取り出した後，時間，色，大きさ，社会性，感情などの表現の男女差を gender preference として抽出している [Liu 07]．また，M. Pasca らはウェブからパターンの一般化を繰り返すことで膨大な事実を取り出している [Pasca 06]．一方，本稿ではナイーブベイズを用いて取り出した男女別・地域別・年代別分類に特徴的な語を手掛かりとして都市風俗を理解することを目指している．

3. ナイーブベイズ分類器

ナイーブベイズ分類器は，クラス c_i $(1 \leq i \leq n_1)$ の事前確率 $P(c_i)$ と素性 $\boldsymbol{x} = (x_1, x_2, \ldots, x_j, \ldots, x_{n_2})$ $(0 \leq j \leq n_2)$ の条件付き確率 $P(\boldsymbol{x}|c_i)$ が与えられたときに，クラスの条件付き確率 $P(c_i|\boldsymbol{x})$ を最大化するクラス $\hat{c}$ を求める問題として定式化され，式 (1) のように表される [元田 06]．

$$\hat{c} = \mathrm{argmax}_{c_i} P(c_i) \prod_j P(x_j|c_i) \quad (1)$$

ここで，ナイーブベイズ分類器の精度を上げるために，タームギャップ指標とクラスギャップ指標を新たに導入する．

タームギャップ指標は分類に効いている素性だけを利用することによって分類精度を高めるための指標であり，$x_j \in \boldsymbol{x}$ なる素性 x_j に対し，その条件付き確率 $P(x_j|c_i)$ を降順に並べたときの隣接する条件付き確率の常用対数の差の最大値をタームギャップ値と定義する．この値が閾値 T_{TG} を越えていればその素性を分類に用いるという操作を繰り返すことにより，分類に効いている素性だけを利用する．

クラスギャップ指標は分類の難しい事例を分類対象から外すことによって分類精度を高めるための指標であり，各事例に対して，クラスの条件付き確率 $P(c_i|\boldsymbol{x})$ を降順に並べたときの上位 2 件の条件付き確率の常用対数の差をクラスギャップ値と定義する．この値が閾値 T_{CG} を越える事例のみ分類対象とすることにより，分類しにくい事例を対象から外す．

4. 男女別・年代別・地域別傾向の分析

4.1 分析データ

分析には Doblog[*1] のデータ[*2] とユーザのデモグラフィック属性を利用した．デモグラフィックデータは，Doblog ユーザを対象に 2005 年 4 月 22 日〜5 月 23 日にかけて行われた「Doblog の利用に関するアンケート」調査結果より得られた性別（男性，女性），年齢（17 才以下，18〜20 才，21〜24 才，25〜29 才，30〜34 才，35〜44 才，45〜54 才，55 才以上），住所（北海道・東北，東京都，東京以外の関東，中部・甲信越，近畿，中国・四国，九州・沖縄，海外）の 3 つの設問結果を利用した．

ブログデータは 758 名のユーザによる約 24 万記事からなるが，1 記事しか投稿していない人もいれば 5272 記事も投稿している人もいてばらつきがある．そこで，100 記事以上投稿している 532 ユーザのそれぞれから 100 記事をランダムに選んで，53200 記事からなるブログコーパスを作成した．作成したコーパスの基本情報を表 1 に示す．

4.2 分類精度

ブログコーパスを MeCab[*3] を用いて形態素解析し，全ての品詞を素性に用いてナイーブベイズによる分類精度を求めた．10 分割交差検定による分類結果を表 2 に示す．表 2 より，タームギャップ，クラスギャップの閾値 T_{TG}，T_{CG} を上げれば Accuracy は上がり Coverage が下がることが分かる．このように Accuracy と Coverage はトレードオフの関係にあるので最適値を決定することは難しいが，本稿では男女分類，年代分類，地域分類のいずれの場合も $T_{TG} = 1$，$T_{CG} = 1$ による学

連絡先: 松村真宏，大阪大学大学院経済学研究科，〒 560-0043 大阪府豊中市待兼山町 1-7，Tel/Fax: 06-6850-5231，matumura@econ.osaka-u.ac.jp

*1　（株）NTT データ，http://www.doblog.com

*2　（株）ホットリンクと（株）NTT データの共同事業契約に基づき，（株）ホットリンクより提供．2003 年 11 月 4 日から 2005 年 6 月 27 日のデータを利用．

*3　http://mecab.sourceforge.net/

表 1: ブログコーパスの基本情報

ユーザの性別	記事数	ユーザの年代	記事数	ユーザの居住地域	記事数
男性	33600	24 歳以下	14600	北海道・東北	4500
女性	19600	25-29 歳	12900	東京都	13400
		30-34 歳	11500	東京以外の関東	16100
		35-44 歳	11000	中部・甲信越	5600
		45 歳以上	3200	近畿	6900
				中国・四国	2700
				九州・沖縄	2700
				海外	1300
計	53200		53200		53200

表 2: 10 分割交差検定による男女分類・地域分類・年代分類の Coverage と Accuracy

閾値		男女分類		地域分類		年代分類	
T_{TG}	T_{CG}	Coverage	Accuracy	Coverage	Accuracy	Coverage	Accuracy
0	0	0.9181	0.5606	0.9262	0.1090	0.9202	0.3703
0	1	0.8084	0.5711	0.7468	0.0925	0.6002	0.4154
1	0	0.7604	0.7927	0.7565	0.3665	0.8351	0.4689
1	**1**	**0.6018**	**0.8564**	**0.2521**	**0.5031**	**0.3286**	**0.6519**
1	2	0.4391	0.8956	0.1334	0.5812	0.1716	0.7508
2	1	0.1842	0.9313	0.0615	0.6759	0.0619	0.7694
2	2	0.1483	0.9507	0.0295	0.8406	0.0435	0.8804
ベースライン		1.0000	0.6316	1.0000	0.3026	1.0000	0.2744

表 3: 男女分類結果のクロス表

	男性	女性	Accuracy
男性	16893	2874	0.8546
女性	1723	10525	0.8593

表 4: 男女別の特徴的な語

順位	男性	PG	女性	PG
1	僕	0.0824	女	0.0451
2	俺	0.0706	旦那	0.0270
3	競馬	0.0178	かしら	0.0266
4	レース	0.0160	あたし	0.0246
5	着	0.0160	ご飯	0.0242
6	勝利	0.0153	とっても	0.0236
7	勝っ	0.0153	夫	0.0220
8	賞	0.0147	母	0.0219
9	野球	0.0140	娘	0.0212
10	走	0.0117	苦笑	0.0192

習モデルを構築した．この学習モデルは，事例数の最も多いクラスを選ぶベースラインと比べると，男女分類で約 22%，地域分類で約 20%，年代分類で約 38%Accuracy が向上している．

4.3 特徴的な語の抽出

学習モデルにおける語の条件付き確率を利用すれば，クラスごとに特徴的な語が抽出できる．まず，語の条件付き確率 $P(x_j|c_i)$ を降順に並べたときの上位 2 件の条件付き確率の差を PG (Probability Gap) 値として求める．この PG 値を用いて語を降順に並べたときに上位にくる語が，そのクラスに特徴的な語となる．

4.4 男女別傾向の分析

男女分類結果のクロス表を表 3 に示す．これより，男性，女性ともに同程度の精度で分類できていることが分かる．

次に，男性・女性に特徴的な語を表 4 に示す．男性は「僕」「俺」など自分について語るのに対し，女性は「旦那」「あたし」「夫」「母」「娘」など家族について語っていることが分かる．また，男性は「競馬」「野球」など娯楽が話題に上るのに対し，女性は「ご飯」など，やはり家庭に目が向いていることが見えてくる．

4.5 年代別傾向の分析

年代分類結果のクロス表を表 5 に示す．表 5 より，24 歳以下と 45 歳以上の分類精度が高く，特に 45 歳以上の Accuracy は約 0.9 もある．

次に，年代別の特徴的な語を表 6 に示す．24 歳以下は「学校」「バイト」「友達」「授業」「テスト」など学校生活に関係する語が多い．また，45 歳以上は「花」「珈琲」「演奏」「夕食」など人生にゆとりが感じられる語が現れている．一方，25 歳〜44 歳で用いられる語は PG 値が低く，この年代には目立った特徴語がないことが分かる．この年代の分類精度が低いのも目立った特徴がないためであろう．

以上より，人生は大きく分けると，24 歳以下の学生モード，25 歳〜44 歳の定常モード，45 歳以上のゆとり探求モードの 3 つのモードがあることが分かる．

4.6 地域別傾向の分析

地域分類結果のクロス表を表 7 に示す．表 7 より,「東京都」および「東京以外の関東」の分類精度が著しく悪く,「海外」の分類精度は極めて高いことが分かる．

次に，地域別の特徴的な語を表 8 に示す.「札幌」「青森」「塩竈」「新宿」「姫路」「名古屋」などの地名や「温泉」「歌舞伎」「ジェフ」など地域に根ざした話題に関する語が現れている．分類精度が約 97% と極めて高い「海外」に関しても,「日本」

表 5: 年代分類結果のクロス表

	24 歳以下	25-29 歳	30-34 歳	35-44 歳	45 歳以上	Accuracy
24 歳以下	4229	295	298	167	454	0.7770
25-29 歳	658	2005	402	198	508	0.5317
30-34 歳	435	313	2023	207	456	0.5891
35-44 歳	319	242	339	1710	632	0.5275
45 歳以上	42	30	40	50	1429	0.8982

表 6: 年代別の特徴的な語

順位	24 歳以下	PG	25-29 歳	PG	30-34 歳	PG	35-44 歳	PG	45 歳以上	PG
1	学校	0.0472	ちなみに	0.0087	旦那	0.0181	つづく	0.0157	花	0.0478
2	バイト	0.0383	なー	0.0082	ちゃっ	0.0145	龍	0.0119	珈琲	0.0367
3	まぁ	0.0304	ダンナ	0.0082	おすすめ	0.0113	依	0.0115	総	0.0282
4	友達	0.0304	ひな	0.0060	苦笑	0.0111	遭遇	0.0111	録音	0.0269
5	授業	0.0298	山尾	0.0054	塩竈	0.0066	羽	0.0110	演奏	0.0267
6	テスト	0.0228	れ	0.0052	小僧	0.0065	鴇	0.0101	夕食	0.0264
7	んで	0.0210	アミ	0.0051	夫婦	0.0060	公邸	0.0089	アクセス	0.0245
8	なんか	0.0201	オイラ	0.0049	本家	0.0058	主人	0.0084	ゆく	0.0245
9	試験	0.0195	バディ	0.0043	ひとこと	0.0056	総理	0.0070	盤	0.0234
10	部活	0.0104	其の	0.0041	せんせい	0.0055	かれ	0.0069	障害	0.0217

表 7: 地域分類結果のクロス表

	北海道・東北	東京都	東京以外の関東	中部・甲信越	近畿	中国・四国	九州・沖縄	海外	Accuracy
北海道・東北	691	42	31	20	70	25	57	194	0.6115
東京都	116	797	119	91	311	175	159	1064	0.2814
東京以外の関東	191	174	1178	217	287	138	173	992	0.3516
中部・甲信越	40	41	62	867	113	65	45	214	0.5992
近畿	56	90	54	57	1410	80	65	459	0.6209
中国・四国	15	25	19	21	101	340	20	115	0.5183
九州・沖縄	20	24	13	14	44	23	673	98	0.7404
海外	1	5	4	1	2	2	10	792	0.9694

表 8: 地域別の特徴的な語

順位	北海道・東北	PG	東京都	PG	東京以外の関東	PG	中部・甲信越	PG
1	温泉	0.0317	新宿	0.0112	こんち	0.0074	旦那	0.0406
2	札幌	0.0215	公邸	0.0076	セガ	0.0066	珈琲	0.0205
3	つづく	0.0214	官邸	0.0058	（株）	0.0065	名古屋	0.0204
4	青森	0.0200	総理	0.0054	許諾	0.0057	俺	0.0202
5	塩竈	0.0191	舞	0.0053	ジェフ	0.0055	あわ	0.0198
6	湯	0.0173	かれ	0.0052	有っ	0.0052	ひな	0.0171
7	せんせい	0.0156	官房	0.0046	よもやま	0.0051	格	0.0165
8	ネ	0.0154	ふる	0.0038	追伸	0.0051	キト	0.0160
9	ひとこと	0.0151	歌舞伎	0.0038	備考	0.0048	岐阜	0.0135
順位	近畿	PG	中国・四国	PG	九州・沖縄	PG	海外	PG
1	や	0.0773	演奏	0.0351	遭遇	0.0463	日本	0.1873
2	へん	0.0281	録音	0.0337	星	0.0456	中国	0.1036
3	やん	0.0256	盤	0.0303	左足	0.0454	日本語	0.0650
4	やろ	0.0217	指揮	0.0279	依	0.0450	アメリカ	0.0589
5	関西	0.0207	楽章	0.0255	匹	0.0449	元	0.0523
6	ねん	0.0193	ＣＤ	0.0232	鴇	0.0448	語	0.0476
7	ええ	0.0160	テンポ	0.0171	龍	0.0436	日本人	0.0467
8	リス	0.0139	広島	0.0169	福岡	0.0302	中国人	0.0400
9	阪神	0.0137	クラシック	0.0169	難	0.0289	諜報	0.0392
10	保護	0.0096	終	0.0169	入手	0.0265	台湾	0.0360

「中国」「アメリカ」「台湾」など海外居住者ならではの視点が特徴的な語として得られている．また，近畿地域は特徴的であり,「や」「へん」「やん」「やろ」「ねん」など関西弁でよく使われる語尾が特徴語の上位を独占している．他の地域の特徴語にはこのような方言が現れていないことから，一般に方言は口頭で用いてもブログ記事を書くときには用いないと考えられる．したがって,「近畿」居住者はブログ記事であっても「言文一致」を貫いていることは大変興味深い傾向であろう．よく関西人は東京に出ても関西弁を使い続けると言われているが，同じことはブログを書く際にも言えるのである．

なお,「東京」および「東京以外の関東」に特徴的な語の PG 値は，他の地域と比べると相対的に小さい．これは「東京」および「東京以外の関東」居住者が使用する語があまり偏っていないことを表している.「東京」および「東京以外の関東」の情報はテレビなどを通して日本全国に発信されることが多いために，この地域の独自性が薄れてしまうのであろう.「東京」および「東京以外の関東」の分類精度が著しく低かったのも，このことが原因だと考えられる．地域分類全体の Accuracy は 0.5031 であるが,「東京」および「東京以外の関東」を除くと Accyracy は 0.6602 に大幅に上がる．

5. まとめ

本稿ではナイーブベイズを使ってユーザの男女別・年代別・地域別傾向を明らかにする方法を提案した．本稿での分析によって得られた知見をまとめると以下のようになる．

- 男性は自分について語り，女性は家庭について語る．
- 人生には，24 歳以下の学生モード，25 歳〜44 歳の定常モード，45 歳以上のゆとり探求モードの 3 つのモードがある．
- 「近畿」居住者は，ブログを書く際にも言文一致を貫く．
- 「東京」および「東京以外の関東」居住者の扱う話題は全国共通．

今回の分析では特徴的な語から考察を進めたが，語の用いられる文脈を把握したより深い分析を行うためには述語項構造を見た方がよい．そのために，ブログコーパスに河原らの手法 [河原 02] を適用して，ブログ格フレーム辞書を構築している．今後は特徴的な語をブログ格フレーム辞書に対応させ，人々の日常を表出化させることを試みる予定である．

謝辞

Doblog の記事データおよび「Doblog の利用に関するアンケート」の調査データは株式会社 NTT データおよび株式会社ホットリンクより提供を受けました．記して感謝致します．

参考文献

[河原 02] 河原大輔，黒橋禎夫：用言と直前の格要素の組を単位とする格フレームの自動構築, 自然言語処理, Vol.9, No.1, pp.3–19 (2002)

[Liu 07] Hugo Liu and Rada Mihalcea: Of Men, Women, and Computers: Data-Driven Gender Modeling for Improved User Interfaces, ICWSM2007 (2007)

[野呂 06] 野呂太一, 乾孝司, 高村大也, 奥村学：イベントの生起時間帯判定，言語処理学会 第 12 回年次大会, pp.837-840 (2006)

[松村 06] 松村真宏，三浦麻子：Doblog の利用に関するアンケート調査からみたユーザ像，第 20 回人工知能学会全国大会, 3D3-3 (2006)

[元田 06] 元田浩，津本周作，山口高平，沼尾正行（著）：データマイニングの基礎，オーム社 (2006)

[Pasca 06] Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, Alpa Jain: Organizing and Searching the World Wide Web of Facts - Step One: The One-Million Fact Extraction Challenge, AAAI (2006)