3B4-NFC-05b-3

# 情報推薦のための Twitter ユーザの性格分析手法

## Personality Analysis Method of Twitter User for Information Recommendation

田中 聡 Satoshi Tanaka　松本 和幸 Kazuyuki Matsumoto　吉田 稔 Minoru Yoshida　北 研二 Kenji Kita

徳島大学大学院 先端技術科学教育学部
Tokushima University, Graduate Schools, Advanced Technology and Sciences

If we can know the personality of users from their tweet messages by analyzing the message based on natural language processing techniques, we think that the personality of users and their preferences can be related each other. In this paper, we created an association between the three levels of five factors expressing personality and Twitter users, and tried to automatically classify the personality patterns by using a Naïve Bayes Classifier. Finally, we analyzed how the features included in their tweets affected each factor.

## 1. はじめに

インターネット技術や情報通信端末の発達および普及により，人々の人格形成および人格発現の場が個人の意見を発現可能な www 上のソーシャルネットワーク・サービス（SNS）へと変化してきている. SNS 上で表現される個人の性格を推定できれば，情報推薦やマーケティング調査などへの応用が可能と考えられる.

言語から性格を分析しようとした研究として，Twitter を利用した研究［1, 2］や，ブログを対象とした研究［3, 4］がある. また，［1, 3］の研究では，単語を感情カテゴリなどに変換することにより分析しているが，Twitter では，個人の趣味の話や砕けた表現を含むことが多く，カテゴリに変換しづらい語彙については無視されてしまう側面がある. 本研究では，素性を名詞に絞り，多様な語彙を扱うことにより，ある性格パターンにおいて共通する語彙を特定することで，カテゴリ化することで欠落してしまう情報を有効に使用することを検討する.

## 2. 関連研究

本研究と類似する関連研究を 2 件紹介する.

### 2.1 筆者の性格推定のための言語調査

那須川ら [1]は，性格に関連する日本語における筆者の言語的特徴を，LIWC(Linguistic Inquiry and Word Count)という単語をカテゴリ化したものを用いて分析している. この研究では，英語で構築された LIWC を日本語化したものを用いて日本の Twitter のユーザの投稿から 79 種類のカテゴリに対応する素性を抽出し，性格プロファイルとカテゴリとの相関に着目して分析をおこなっている.

### 2.2 Big Five を用いたブログ著者の性格推定

奥村ら[3]は，感情判断システムを用いて，ブログの印象から推定される擬似的な性格と，NEO-FFI 検査との結果との比較をおこなっている.

### 2.3 本研究との差異

本研究と関連研究との差異を以下に示す. 本研究ではツイートから名詞を取得しそのまま使用しているが，那須川らの研究では LIWC を使用することによって日本語をカテゴリ化し，全 79 のカテゴリを定義し，各カテゴリに該当する表現をテキスト中から特定する. また，素性として名詞だけではなく動詞や形容詞なども素性として採用している点が異なる.

奥村らの研究との違いとしては，感情判断を用いていることと，本研究ではエゴグラムを使用しているのに対して，奥村らの研究では Big Five を使用している点があげられる.

連絡先： 松本 和幸，徳島大学理工学部，徳島県徳島市南常三島町 2-1，matumoto@is.tokushima-u.ac.jp

## 3. 提案手法

本研究では，Twitter API を用いてエゴグラムの全 243 パターンに可能な限り対応できるよう複数のパターンのエゴグラム回答者のツイートを取得し，それを形態素解析エンジン MeCab [5]で形態素解析し，名詞を抽出したものを教師データとする. さらにこのデータを用いて任意のユーザがどのエゴグラムのパターンに最も当てはまるかを，ナイーブベイズ分類器を用いて自動分類する. さらに分類結果より，どのような発言（形態素）が性格を推定する際に影響したかを分析し考察する. 以下，それぞれについて詳述する.

### 3.1 ツイートの収集

Twitter API を用いて，ナイーブベイズ法による性格パターン分類器の構築に必要なツイートのデータを取得する. ただし，エゴグラム診断サイト「エゴグラムによる性格診断」[6]にてエゴグラムの診断をおこない，その結果をツイートしたユーザのみを収集対象とし，そのアカウントがエゴグラム診断の結果をツイートした日時以降のツイートを 200 件ほど取得した.

### 3.2 ツイートにおける素性の抽出

取得したツイートデータに対して形態素解析をおこない，ツイート文から名詞を抽出する.今回名詞を素性とした理由として，品詞ごとに形態素解析をおこなったときの名詞の出現数の多さおよび種類数の多さがあげられる.

また，本研究で名詞以外の品詞を素性として採用しない理由について述べる. まず，どのような語句が性格パターンの分類に影響を与えるかは現時点で不明ではあるが，助詞など，その単語単体で意味を成さない品詞を素性に取り入れると，分類器の精度低下に影響することが考えられる. 形容詞や副詞，動詞などの語句にユーザの個性が表れる可能性はあるが，その種類数から，名詞ほどの差異が出るとは考えにくい. また，素性データ量が膨大となることにより，性格パターン分類結果の人手による分析にかける時間的コストが大きくなると考えたため，今回は名詞のみに着目する.

### 3.3 エゴグラム

本研究における性格推定の根拠となるエゴグラムについて紹介する．エゴグラムとはエリック・バーン博士の交流分析で用いられている P（Parent），A（Adult），C(Child)の 3 つの自我状態を，弟子であるジョン・M・デュセイが P を厳しい親を表す CP（Critical Parent）と優しい親を表す NP（Nurturing Parent）に分け，C を自由奔放な子供を表す FC（Free Child）と従順なこどもを表す AC（Adapted Child）に分け，計 5 つの自我状態をもとに性格を推定する性格診断法である．これらの 5 つの自我状態の高低を図示したグラフをみることで性格や相性を判断する[7, 8].

また，本研究ではエゴグラムを使用しているが，ほかによく用いられる性格分析の手法として，Big Five というものがある．斎藤らの研究[9]では，5 つの特性を用いて個人のパーソナリティ特性を捉えられる Big Five を使用した分析をおこなっている．エゴグラムで用いられる 5 つの自我状態について簡単にまとめたものを表 3 に示す．この表に示すとおり，各自我状態の高低により，性格の傾向を判断することができる．本研究では，各自我状態が高い場合に’a’, 普通の場合に’b’, 低い場合に’c’という記号で表し，パターンを表すときには，’CP’, ‘NP’, ‘A’, ‘FC’, ‘AC’の順に高低を表す記号を並べて’abaac’ のように表すことにする．

表 3: エゴグラムにおける 5 つの自我状態

| 自我状態 | 高い場合 | 低い場合 |
|---|---|---|
| CP（お父さん度） | リーダーシップ | 無責任 |
| NP（お母さん度） | 思いやり | 冷淡 |
| A（大人度） | 冷静 | 感情的 |
| FC（やんちゃ坊主度） | 好奇心 | 無気力 |
| AC（いい子ちゃん度） | 協調性 | 反抗的 |

### 3.4 発言特徴の分析

ここで，前述した分類器によってユーザ 5 人に対して自動分類した結果が以下のようになったとする（表 4）．

表 4：推定結果一覧

| ユーザ | 推定結果 |
|---|---|
| U1 | bbbba |
| U2 | cbaba |
| U3 | bcaca |
| U4 | aacac |

例として，表 2 の上から 3 人までの推定結果に注目すると，AC という順応性を表す 5 番目の因子が a であるという共通点が見つかる．本研究においては，このような各自我状態における共通点（高低の一致）に着目し，また，ユーザの発言において共通に出現する名詞から，なぜ自我状態 AC において a という評価が得られたかについて分析する．

### 3.5 学習および分類の流れ

性格の自動分類のための学習データ作成の流れから性格分類器を用いて分類結果を出力するまでを図 5 に示す．

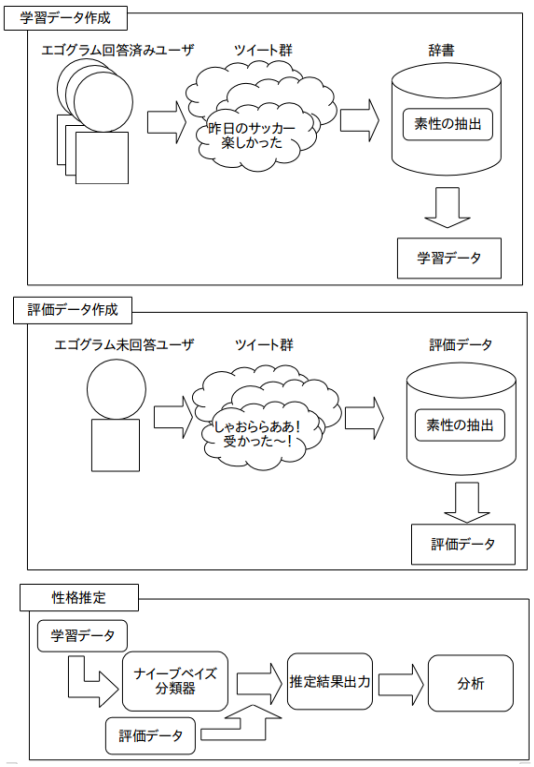


図 5： 学習および分類の流れ

## 4. 実験 1 -性格推定-

### 4.1 実験に用いるデータ

本実験で教師あり学習のデータとして使用するデータは Twitter API を用いて Twitter より取得したものであり，取得する条件としてつぎの 2 つがある．

1. エゴグラム診断サイトでエゴグラムの回答をしており，その結果をツイートしているユーザ
2. bot などのような決められた時間ごとにいくつかの定型文を機械的にツイートするだけのアカウントではないこと

以上の 2 つの条件を満たしているユーザアカウントを対象にパターンごとに 200 ツイートずつ取得した．これらの条件で実験データ収集をおこなった結果，約 200 人分のエゴグラム回答者のデータを収集でき，エゴグラム全 243 パターン中約 100 パターンのデータを取得することができた．残りのパターンについては得ることができなかった．取得したデータの収集数ごとのパターン数についてヒストグラムで表したものを図 6 に示す．

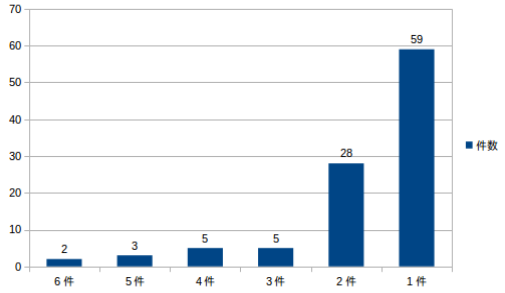


図 6: 収集できたユーザ数ごとのパターン種類数

### 4.2 実験の流れ

本実験の大まかな流れと簡単に図にしたものを以下に示す.

1. 前準備としてエゴグラムのパターンごとにツイートデータを Twitter API を用いて取得
2. 形態素解析器エンジン MeCab を用いて 1 で得たデータを形態素解析し, 素性を抽出
3. 抽出した名詞とエゴグラムの各パターンを対応させ, ユーザごとに教師あり 学習
4. 抽出した特徴語を用いてナイーブベイズ分類器で分類
5. 4 の結果からどのような特徴語が推定に影響を及ぼしたのかを分析

図6で示したようにパターンによっては 6 ユーザ分のデータがあるものや, 1 ユーザ分のデータしかないものがある. 本実験では, 学習の偏りを避けるため, あるパターンにおいて複数ユーザのデータが取得できている場合も, すべてのユーザのデータを使用せず, ランダムで選択した 1 ユーザ分のデータのみを学習データとして使用する.

### 4.3 実験結果

実験における評価用データとして, 学習データと同様の収集方法により取得したデータのうち, 学習データとして選択されなかったユーザのデータを用いる. また, 本研究では発言特徴の分析をすることが最終目標であるが, その過程として, 構築した分類器により出力された性格推定結果も重要であるためそれぞれについてまとめる.

エゴグラムのパターンが重複したため学習データに用いなかったユーザアカウント 30 人分を評価データとして用いて性格推定の評価をおこなった. 評価指標は, 推定精度(一致率)を用い, 式1により計算する.

$$\text{推定精度(一致率)[\%]} = \frac{\text{各自我状態一致数} * 100}{5} \quad (1)$$

構築した性格パターン分類器を用いて評価用データを自動分類した結果, 正解パターンとの平均一致率は約 47%となった. 結果の一部を表 7 に示す. 学習に用いたパターン数が膨大であるため, 完全一致はみられなかったが, 60%の一致率(3 つの自我状態の高低が一致)という高い一致率が得られるパターンも存在した.

表 7: 分類結果(一致率)

| 正解パターン | 分類パターン | 一致率 |
|---|---|---|
| bbbab | bbabb | 60% |
| cabca | bbbaa | 40% |
| bbbbb | cabbb | 60% |
| bbbba | baabb | 40% |
| bbabb | cabbb | 40% |
| babab | bbabb | 40% |
| aabab | abacb | 40% |
| aaabb | babaa | 20% |
| bbaba | aabba | 40% |

### 4.4 比較実験

比較実験として, エゴグラムの各自我状態ごとに高低をラベル付けして学習した性格パターン分類器を自我状態の数だけ用いる手法について実験し, 評価をおこなった. 4.3 で使用したものと同じ評価用データを分類した結果を表 8 に示す.

表 8: 比較実験における分類結果(一致率)

| 正解パターン | 分類パターン | 一致率 |
|---|---|---|
| bbbab | bbaaa | 60% |
| cabca | baaaa | 40% |
| bbbbb | bbaaa | 40% |
| bbbba | baaaa | 40% |
| bbabb | bbaaa | 60% |
| babab | bbaaa | 40% |
| aabab | babaa | 60% |
| aaabb | bbaab | 40% |
| bbaba | bbaaa | 80% |

平均一致率として約 51%という結果が得られた. 以上より, 自我状態ごとに分類器を用いる場合のほうが分類性能は優れているといえる.

## 5. 実験 2 -性格分析-

### 5.1 分析結果

エゴグラムのそれぞれの自我状態ごとに分類結果と, そのユーザにおける出現素性およびツイート例を表 9～表 18 にまとめ, そのパターンに分類された要因について分析し, 考察する. また, 本節では, 分析において各自我状態が a の場合に「高い」, 低い場合に「c」とする.

・CP(お父さん度)
高い＝リーダーシップ, 他人に対して批判的　低い＝無責任

表 9: CP の高いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| バカ, アホ, クソ | まぁクソはどんな政治形態で政治を行ってもクソ |

表 10: CP の低いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 日, 今日 | 今日はおなかいっぱい食べてしまった(罪悪感) |

高いユーザでは、「バカ, アホ」などの他人に対して批判的な面が見られ、低いユーザでは, 無責任という特徴に当てはまる素性は見られなかった.

・NP(お母さん度)
高い＝思いやり　低い＝冷淡

表 11: NP の高いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 人 | 今この人出てるなら見るって人だけ確認してた |
| みんな | 正月でみてないみんなは見てね |

表 12: NP の低いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 嫌い | (′ ;ω;`)茄子嫌い |

高いユーザでは「人, みんな」から他人を気にかける面が見られ, 思いやりの特徴に当てはまるといえる.低いユーザでは「嫌

い」などネガティブで後ろ向きな面が目立ち，冷淡とまではいえないが方向性としては似ているといえる．

・A（大人度）

高い＝冷静，安定　低い＝感情的

表 13:　A の高いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 幸，幸せ | …互いに支えあう幸せな家庭を維持し |

表 14:　A の低いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 金，課金 | ね!！課金は世界を動かす |

高いユーザでは「幸せ」などの素性から幸せを求める安定思考の人間である人間といえ，安定の特徴に当てはまるといえる．低いユーザでは「課金」などの素性から我慢強さなどがなく衝動的かつ感情的だといえる．

・FC（やんちゃ坊主度）

高い＝好奇心　低い＝無気力

表 15:　FC の高いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 好き，楽 | 好きなバンドマンが… |

表 16:　FC の低いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 室，室内，中 | すぐ室内温度が 20℃を… |

高いユーザの「好き，楽」などからは好奇心などの特徴は見られないが、低いユーザでは「室，室内」から室内の話題が多く内向的な面が見られ，無気力とはいわないがそれに近いものがあるといえる．

・AC（いい子ちゃん度）

高い＝協調性　低い＝反抗的

表 17:　AC の高いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| 感，身体 | 栄養バランス考えられてる感がある。。。 |

表 18:　AC の低いユーザ

| 共通した素性 | ツイート例 |
|---|---|
| アホ，バカ | 素晴らしいアホの連鎖や |

高いユーザからは、協調性に関する特徴は見られなかったが，低いユーザでは「アホ」などの他人を見下すような面が見られた．

### 5.2 分析結果の考察

ユーザの性格と関連の強いであろう素性を，自我状態の高い・低いユーザの発言に共通して出現する名詞を分析することで調べてみた結果，関連が無いとはいえないが，あまり強い関連を見出すことができないパターンも多かった．この原因として，パターンごとに収集できたユーザ数が少ないことと，LIWC のようなカテゴリ化をしなかったために，共通する語句（同義・類義語の判断も含め）を目視により判別することになり，分析が十分でなかったことがあげられる．しかし，今回の分析を通して，名詞が素性として性格パターンの分類に役立つ要素であることがわかった．ほかの品詞（動詞，形容詞，副詞，etc.）についても同様に分析していくことにより，性格パターンの分類には，学習データがどの程度のユーザ数分必要かを見積もっていく必要があると考える．また，今回は性格が一定の場合を想定したが，ツイートは時系列で分析可能であるため，一定期間ごとに性格推定をおこなうことによる，性格の変動についても分析が必要であると考える．

## 6 おわりに

本研究では，Twitter のツイートからエゴグラムを推定し，そのツイートの名詞からどのような名詞が性格推定に影響を与えているかを分析した．その結果，一定の推定精度が得られ，ツイート内容からの性格推定が可能であることが分かった．また，分析結果から，共通名詞を調べることによって，各自我状態のレベルを特定する際に，名詞が有用な素性であることを明らかにした．

今後の課題として，エゴグラムの 243 パターンのなかで集めきれなかったパターンについて収集をし，各パターンについてそれぞれユーザ数を増やした状態での評価実験をおこない，さらに詳細な分析をおこないたいと考える．さらに，時系列で性格推定し，ユーザの性格と時期・季節・時間帯などとの関連性を探ることで，情報推薦への応用可能性を検討したい．

また，分類に用いる手法としてナイーブベイズ法だけではなく，SVM などの分類手法も用いて，素性についても様々なパターンを考慮し，分類結果の比較をおこない，より適した分類手法を模索したい．

## 謝辞

本研究は JSPS 科研費 15K00425,15K00309,15K16077 の助成を受けたものです．

## 参考文献

[1]. 那須川哲哉，上條浩一，山本 眞大，北村 英哉，“日本語における筆者の性格推定のための言語的特徴の調査”，言語処理学会第 22 回年次大会発表論文集，pp.1181-1184，2016.

[2]. 岡本拓馬，松本和幸，吉田稔，北研二，“ナイーブベイズ法を用いた Twitter による性格推定”，言語処理学会第 20 回年次大会発表論文集，pp1123-1125，2014.

[3]. 奥村紀之，金丸祐亮，奥村学，“感情判断と Big Five を用いたブログ著者の性格推定に関する調査”，人工知能学会全国大会論文集第 29 回，pp.1-4，2015.

[4]. 南川敦宣，横山浩之，“テキストマイニングによる個人 Blog データからの性格分析手法”，データマイニングと統計数理研究会第12回，pp. 96-100，2010.

[5]. MeCab，http://taku910.github.io/mecab/ .

[6]. エゴグラムによる性格診断 http://www.egogram-f.jp/seikaku .

[7]. ジョン・M・デュセイ，“エゴグラム”，創元社，2000.

[8]. 東京大学医学部心療内科 TEG 研究会，“新版 TEGII 解説とエゴグラム・パターン”，金子書房，2006.

[9]. 斎藤崇子，中村知靖，横山まどか，“性格特性用語を用いた Big Five 尺度の標準化”，九州大学心理学研究，pp135-144，2001.