G-09

Twitter におけるユーザの性格推定
Personality estimation of Twitter users

津崎　誠也†　　山本　博史†‡
Seiya Tsuzaki　　Hirofumi Yamamoto

1. 序論

近年，スマートフォンやタブレット端末の情報機器が普及したと同時にネットショッピングやオンラインゲーム等インターネット上におけるサービスも増加した．アップル社が運営する App Store ではモバイル向けアプリの数が 2016 年現在までで約 200 万本リリースされたと言われている．しかし，サービスが急激に増加していることから，ユーザは自分に適するサービスを探すことが困難になってきている．そこで，嗜好と密接な関係性にある性格に着目した．矢澤ら[1]による研究にて，性格と嗜好には密接な関係がある可能性があることがわかる．人間は他者の発話を受け「この人は温厚な人だ」「この人は情緒不安定な人だ」というように，発話した内容から話者の性格を推定する能力を利用して，コミュニケーションの円滑化や類似した性格の者同士によるコミュニティの形成等を行っている傾向にある．そのため，嗜好に応じて類似した人の購入傾向を調査することでターゲティング広告やワン・トゥ・ワン・マーケティングにおいても，性格から嗜好を推定することの有用性は高いと考えられる．ユーザに情報を提供する研究[2]やユーザの性別を判別する研究[3,4]のように Twitter 上の発言内容からユーザに情報を提供する試みやユーザの人物像を推定する試みが行われている．本研究では，機械学習によって Twitter 上におけるユーザの発言内容から性格を推定するシステムの構築を試みる．

2. 先行研究

筆者らが以前行った研究[5]では田中ら[6]の使用したナイーブベイズ分類器を利用して，フィクション上のキャラクタの性格推定を行った．この研究ではフィクション上のキャラクタを対象に学習データを収集して，性格が未知のキャラクタの発言内容から性格を推定している．また，この時収集した学習データは対象のキャラクタの発言内容と Big Five における 5 つの性格要素の分類を行い，それらをキャラクタの性格とした．前研究での実験結果から Big Five を利用したナイーブベイズ法による性格推定は有用であると考えられており，本研究では次段階として実在する人間の発言内容から話者の性格を推定するために Twitter のユーザを対象に学習データを収集し，性格が未知のユーザの発言内容から対象ユーザの性格の推定を試みる．

†近畿大学総合理工学研究科，
Graduate School of Science and Engineering, Kinki University
‡近畿大学理工学部，
Department of Science and Engineering, Kinki University

3. Big Five を用いた性格推定

Big Five とは Lewis R. Goldberg が提唱したパーソナリティの特性論 1 つで，人間が持つさまざまな性格は 5 つの要素の組み合わせで構成されるとするものである．その 5 つの要素を以下に示す．

- 経験への開放性（Openness to Experience）
  好奇心が強い・警戒心が強い
- 勤勉性（Conscientiousness）
  真面目・不真面目
- 外向性（Extroversion）
  社交的・おとなしい
- 協調性（Agreeableness）
  温厚・冷徹
- 情緒不安定性（Neuroticism）
  神経質・自信家

本研究ではユーザの各要素の尺度を測定するために，従来使用されている質問用紙を利用してユーザの性格を調査する．質問用紙による質問内容は[7]を参考にした．そして，ユーザの性格を 5 組の相対する各 2 種合計 10 種の性格にそれぞれ分類する．

4. ユーザの特徴量の抽出

4.1　Twitter を用いた発言内容の取得

ユーザの性格を表す特徴量を抽出するために必要となる学習データを Twitter API を用いて取得する．3 章にて述べたアンケートを行ったユーザを対象に最新ツイートを取得する．また，本研究では前研究と違い，ユーザ本人の意図により文章が入力されるため，句読点や記号等の単語に関しても学習データとして取得する．取得する際，URL 等のユーザが意図しない単語が含まれた発言や画像が添付された発言は発言者の性格に依存しないものと考えられるため，収集しないものとする．

4.2　ユーザの特徴量の抽出

本研究では推定に使用する特徴量として，ユーザの発言内容から得られる形態素を使用する．各形態素のエントロピーを求め，それぞれの性格に頻出する形態素のうち，エントロピーの絶対値の差が大きい形態素を性格推定に適した特徴量として扱う．各組の性格中に存在するそのエントロピーの絶対値の差が大きい同一形態素のエントロピーの絶対値を求める．そして，同一形態素の確率比を計算し，どちらによく頻出する単語であるか求め，正負の符号をつける．特徴量を選択する際，正側の特徴量のエントロピーの絶対値の差の合計値と負側の特徴量のエントロピーの絶対値の差の合計値が 0 に近づくように選択することで偏りが発生しないようにする．

4.3 判別分析方法

本研究では 2 章で述べた田中らの研究にて使用されたナイーブベイズ分類器を用いて性格の推定を行う．推定対象のユーザに出現する単語の出現回数並びに出現確率を用いて特徴量の尤度合計でどちらの性格グループに分類されるか判別する．また，本研究では収集した学習データ内の性格に偏りがあるため，事前確率を用意することで偏りを解消する．

5. 性格推定実験

5.1 実験手順

3 章にて述べたユーザを対象に発言内容を収集したところ，総ユーザ数 41 人，約 3,000 ツイート，全形態素数 937,557 単語を収集し，これらを性格推定の為の学習データとする．収集した形態素数の平均は約 21,633 単語，標準偏差は約 11,911 であった．次に 3 章にて説明した方法で Big Five の要素ごとに対象のユーザを分類したところ，表 1 のように分類された．

表 1 Big Five 各要素の分類

Big Five の要素	高得点(人)	低得点(人)
経験への開放性	26	14
勤勉性	19	20
外向性	28	12
協調性	33	7
情緒不安定性	31	8

次に性格推定に必要な特徴量の抽出を行う．本研究では形態素解析ツールとして Chasen[8]を用いた．なお，品詞情報は今回使用していない．次に，1-gram モデルを作成する．1-gram モデルとは形態素の総数を母数として，各形態素の出現頻度を確率で表したモデルである．作成した 1-gram モデルから確率値を算出し，4.2 章にて説明した方法で各形態素のエントロピーを算出する．そして，上位数 10 の形態素を本実験にて使用する特徴量とする．

5.2 使用した特徴量の例

前章にて説明した手順を踏まえて各特徴量の 1-gram を算出した結果，各要素中の性格推定のための特徴量として使用した形態素を下記の表に示す．

表 2 経験への開放性の推定に使用した特徴量と
エントロピー（一部抜粋）

特徴量(形態素)	エントロピーの差	正負
。	0.011639336	正
・	0.011271298	正
!	0.021819085	負
w	0.014219536	負

表 3 勤勉性の推定に使用した特徴量と
エントロピー（一部抜粋）

特徴量(形態素)	エントロピーの差	正負
。	0.022262899	正
w	0.012972483	正
(	0.009353136	負
笑	0.009353136	負

表 4 外向性の推定に使用した特徴量と
エントロピー（一部抜粋）

特徴量(形態素)	エントロピーの差	正負
。	0.012919467	正
…	0.009784212	正
w	0.014023437	負
!	0.012177251	負

表 5 協調性の推定に使用した特徴量と
エントロピー（一部抜粋）

特徴量(形態素)	エントロピーの差	正負
…	0.015662753	正
笑	0.010764223	正
!	0.011343046	負
w	0.011315764	負

表 6 情緒不安定性の推定に使用した特徴量と
エントロピー（一部抜粋）

特徴量(形態素)	エントロピーの差	正負
…	0.009079802	正
w	0.007134211	正
!	0.012373877	負
。	0.005841445	負

4.2 章で説明した手順を踏まえ，特徴量を抽出したところ，
経験への開放性における特徴量は 28 種うち好奇心の強い性格によく見られる特徴量から 10 種，警戒心の強い性格によく見られる特徴量から 18 種抽出した．勤勉性における特徴量は 27 種うち真面目な性格によく見られる特徴量から 10 種，不真面目な性格によく見られる特徴量から 17 種抽出した．外向性における特徴量は 22 種うち社交的な性格によく見られる特徴量から 10 種，おとなしい性格によく見られる特徴量から 10 種抽出した．協調性における特徴量は 29 種うち温厚な性格によく見られる特徴量から 10 種，冷徹な性格によく見られる性格から 19 種抽出した．情緒不安定性における特徴量は 26 種うち神経質な性格によく見られる性格から 10 種，自信家な性格によく見られる性格から 16 種抽出した．

5.3 性格推定実験

新たなユーザ 10 人の発言内容を入力データとし，性格推定実験を行う．これらの入力データは 3 章にて説明した学習データと同様にグループを分け，形態素解析等の処理を行い，5.1 章にて定めた本研究にて使用する特徴量のエントロピーから正しいグループに分けられるか実験する．

10 人のユーザを Big Five の要素ごとに分類したところ，表のように分類された．

表 1 Big Five 各要素の分類

Big Five の要素	高得点(人)	低得点(人)
経験への開放性	8	2
勤勉性	4	6
外向性	7	3
協調性	8	2
情緒不安定性	9	1

5.4 実験結果

5.3 章にて説明した実験結果についてまとめたものを表 6 に示す.

表 6 実験結果

Big Five の要素	正答数	正答率
経験への開放性	8	80%
勤勉性	7	70%
外向性	4	40%
協調性	3	30%
情緒不安定性	9	90%
合計	31	62%

経験への開放性及び情緒不安定性においてはそれぞれ 80 パーセント，90 パーセントと正答率が高く，外向性においては 40 パーセント，協調性においては 30 パーセントと正答率が低かった．二項検定においても，経験への開放性及び情緒不安定性においてはそれぞれ p=0.0439 p=0.0094 と片側検定で 5 パーセントの有意差がみられた．しかし，外向性及び協調性においてはそれぞれ p= 0.205，p= 0.115 と有意差がみられなかった．また Big Five 全体における正答率は 62 パーセント，二項検定では p= 0.0270 と片側検定で 5 パーセントの有意差が見られた.

6. 考察

今回行った実験から Big Five を使用した本手法は実在する人物に対して有用であると考えられる．本研究は前研究と全体を比較すると有意である見込みがあるが，各要素の比較を行うと前研究よりも正答率が低下している要素もあることから学習データに大きく依存していることがわかる．これは小比田ら[9]の研究にて Twitter 等 SNS 上での CMC(Computer-Mediated Communication)におけるシャイな人々は承認欲求を満たしたい傾向にあると示唆していることから，おとなしい人や温厚な人が Twitter では承認欲求を満たすために普段は使用しない単語を使用している可能性があるため，外向性と協調性においては分類が困難であると推測される.

今回は品詞情等を考慮せず，特徴量を選択したため，性格推定に使用した特徴量は句読点や記号等が多く，何の文の 1 文なのか判別がつかない形態素が多かった．これらの形態素が性格のイメージに関しているとは思い難い.
また，今回は実験データが少なかったことから，勤勉性の正答率は高かったが，有意差が見られなかったと考えられる.

7. 今後の課題

今後の課題としては，学習データの大幅な増量や N≧2 以上の N-gram モデルを用いた実験，品詞情報を考慮したアプローチが挙げられる．特に，品詞情報を考慮したアプローチに関して，今回 1-gram モデルにて特徴量を選択した結果，各要素を共通して「。」「…」「(」等の記号が多く選択された．「。」や「…」等の場合，句点としてではなく，顔文字やアスキーアートと呼ばれるプレーンテキストによる視覚的表現技法に用いられており，実際の意味と異なる場合に使用する箇所が多く存在した．この問題を解消するために N≧2 以上の N-gram モデルを用いることで，実際の意味で使用されている場合と異なった意味で使用されている場合と区別して，特徴量を選択できると推測できる.

特徴量を選択する際，正側の特徴量のエントロピーの絶対値の差の合計と負側の特徴量のエントロピーの絶対値の差の合計値が 0 に近づくように選択することで偏りが発生しないようにしているが，実数値にて実験を行った場合，0 になることはなく，偏りが発生した．この偏りが本研究にて影響があるのか各要素の合計値の差における特徴量の抽出方法を再検討する必要があると考えられる.

本論文における実験にて正答率は高かったが有意差が見られなかった勤勉性は，実験データをさらに多く用意して実験を行うことで問題を解消することができると推測する.

参考文献

[1]矢澤 櫻子，星野 准一，宇津呂 武仁：”性格分析・色嗜好に基づく高解像度情報推薦に関する考察”，情報処理学会,Vol.2016-HCI-167, No5，pp.2-7，2016

[2]田中 聡，松本 和幸，吉田 稔，北 研二：”情報推薦のための Twitter ユーザの性格分析手法”， 人工知能学会全国大会，pp.1-4，2016

[3] 長浜 祐貴，遠藤 聡志，當間 愛晃，赤嶺 有平，山田 考治：”ユーザーツイート解析による人物像推定手法の提案と検討”，第 76 回全国大会，2014

[4] 佐古 龍，原 元司：”Twitter 利用者の性別判定システムの構築”，第 29 回ファジィシステムシンポジウム，2013

[5] 津崎 誠也，山本 博史：”キャラクタの性格推定”，2017 年度情報処理学会関西支部 支部大会，2017

[6] 田中 翔一，山本 博史：”フィクション上のキャラクターに対する性格推定”，言語処理学会 第 78 回全国大会，pp.597-598，2016

[7] 堀洋道,”心理尺度集 I”,山本眞理子(編),サイエンス社,東京,2001

[8] Chasen，http://chasen-legacy.osdn.jp/，version2.3.3

[9] 小比田 涼介，宮本 エジソン 正：” Twitterでのシャイな人々の自己開示～行動シャイネスと自己開示抑制～”，日本認知科学会 第32回大会，pp.2-23，2015