

属性が性格に及ぼす作用の存在解明のための ツイッターユーザの属性抽出と性格ベクトル化

石川 晴貴[†] 荒澤 孔明[†] 服部 峻^{††}

^{†,††}室蘭工業大学 ウェブ知能時空間研究室 〒050-8585 北海道室蘭市水元町 27-1

E-mail: [†]{16024011,18096001}@mmm.muroran-it.ac.jp, ^{††}hattori@csse.muroran-it.ac.jp

あらまし 我々人間には性別、人種や宗教など様々な属性が存在する。このような属性は少なからずその属性の所有者である我々に対し、影響を与えているのではないであろうか。例えば、自分がスポーツを趣味としているのならば、その趣味によって何らかの性格に変化がもたらされているのではないかということである。仮に属性がその所有者の性格に作用しているのならば、人間に新しく属性を付与することでその性格を変化させられる可能性がある。また、その属性が性格に対し、どういった作用を及ぼしているのかについて解明することができれば、目的に沿って、例えば、うつ病や躁うつ病などの気分障害の予防や症状の緩和などにも応用できるのではないかと考える。そこで本稿では、SNSの分析を用いて、属性が性格に及ぼす作用の存在解明を行うことを目的としたツイッターユーザの属性抽出と性格ベクトル化の手法を提案する。

キーワード 属性抽出, 性格ベクトル化, SNS分析, 機械学習

Property Extraction and Personality Vectorization of Twitter Users for Clarifying the Existence of Property's Effects on Personality

Haruki ISHIKAWA[†], Komei ARASAWA[†], and Shun HATTORI^{††}

^{†,††} Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27-1 Mizumoto-cho, Muroran, Hokkaido 050-8585, Japan

E-mail: [†]{16024011,18096001}@mmm.muroran-it.ac.jp, ^{††}hattori@csse.muroran-it.ac.jp

Abstract We, humans, have various properties such as gender, race, and religion. Such a property could have an influence on us who are the owners of those properties. For example, if you have sports as one of your hobby, then the hobby may have changed your personality. If a property affects the personality of the owner, it may be possible to change the personality by assigning a new property to a person. In addition, if it is possible to elucidate what kinds of effects the property has on the personality, it may be applicable to prevention of mood disorders such as depression and manic depression. Therefore, this paper proposes a method for property extraction and personality vectorization of Twitter users for the purpose of clarifying the existence of the effects of properties on personality using SNS analysis.

Key words Property Extraction, Personality Vectorization, SNS Analysis, Machine Learning

1. はじめに

我々人間には性別、人種や宗教など様々な属性が存在する。こういった属性は少なからずその属性の所有者である我々に対し、影響を与えているのではないであろうか。例えば、自身がサッカーを趣味としているのならば、同じサッカーを行う仲間たちやその周囲の環境から影響を受け、その性格が活発になるなど何らかの変化がもたらされているのではないかということ

ある。仮に属性がその所有者の性格に作用しているのならば、人間に新しく属性を付与したり、逆に取り除いたりすることで、その性格を変化させられる可能性がある。また、その属性が性格に対し、どういった作用を及ぼしているのかについて解明することができれば、目的に沿って、例えば、うつ病などの気分障害の予防や症状の緩和などにも応用できる可能性がある。さらに詳しく言えば、あるうつ病の人間の属性と性格ベクトルを分析することで、その人にどのような属性を新たに付与すれば

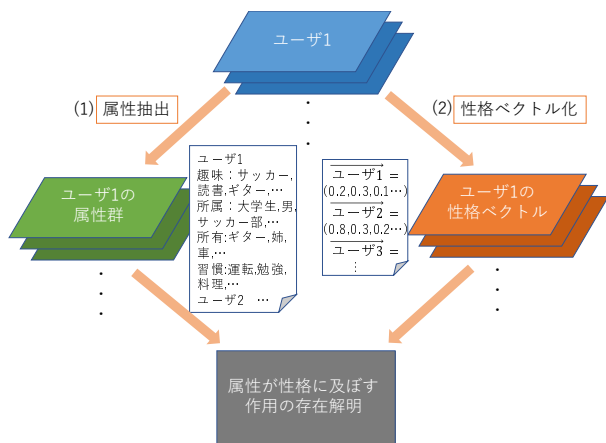


図 1 本研究の概観

症状が改善できるかや、今所持しているどの属性の中のうち病の要因があるのかを調べることができる可能性がある。

このような属性が性格に及ぼす作用の存在解明のためには、図 1 の (1) と (2) のように、大量の人間 (の SNS やライフログなど個々人のデータ) から属性抽出と性格ベクトル化を行い、これらから属性と性格の間に何らかの関係があることを明らかにした上で、その因果関係を分析する必要がある。そこで本稿では、その第一歩として、ツイッターアカウントを対象とし、属性が性格に及ぼす作用の存在解明を目的とした属性抽出と性格ベクトル化の手法を提案する。なお、性格ベクトル化の手法については 2 章で述べ、3 章以降では、未だ手法が確立されていない、属性が性格に及ぼす作用の存在解明のための属性定義とその抽出手法について述べる。

2. 性格ベクトル化

2.1 性格ベクトル化のタスク

本研究では大量の人間の属性抽出と性格ベクトル化を行い、それらを用いて属性が性格に及ぼす作用の存在解明を行うことを目標としている。そのため、人手による大量の人間の性格ベクトル化作業は困難であり、自動化する必要がある。

2.2 性格ベクトル化の手法

本稿では、2.1 節のタスクに対して、IBM Watson の Personality Insights [1] によって分析されたビッグ・ファイブの数値を利用し、人間の性格をベクトル化する手法を提案する。Personality Insights は文書を入力とし、その文書の書き手のビッグ・ファイブ、価値、ニーズを分析する。ビッグ・ファイブとは外向性 (Extraversion)、協調性 (調和性; Agreeableness)、勤勉性 (誠実性; Conscientiousness)、神経症傾向 (情緒 [不安定性; Neuroticism)、開放性 (経験への開放性; Openness) という 5 つの次元でパーソナリティの全体的構造を捉えるモデルである [2]。Personality Insights ではそれぞれ、外向性、協調性、誠実性、感情起伏、知的好奇心と日本語訳されている。なお、本稿ではビッグ・ファイブの各項目について 0 から 1 の間に正規化された Percentile の値を用いて、5 次元のベクトルを生成する。これにより、大量のユーザの性格ベクトル化をアンケートなどの人手を媒介せずに行うことが可能となる。

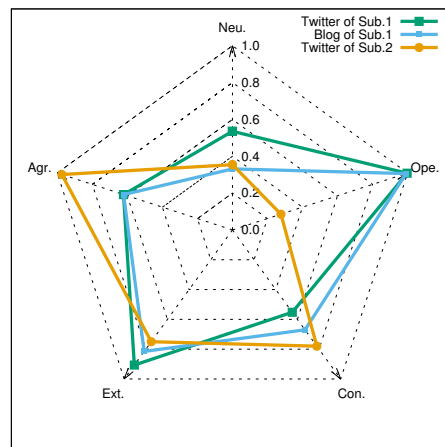


図 2 Personality Insights によって分析されたビッグ・ファイブ

2.3 本研究への適用可能性の検討

本節では、ある被験者 1 のツイッターとブログでの投稿と、ある被験者 2 のツイッターでの投稿を Personality Insights を用いて分析し、それらの結果から本研究への Personality Insights の適用可能性を検討する。

図 2 は、実際に被験者 1 のツイート (約 2,300 件)、ブログ記事 (約 200 件)、被験者 2 のツイート (約 2,800 件) の 3 種類の文書をそれぞれ Personality Insights にて分析し、5 次元のレーダーチャート上にプロットした結果である。

被験者 1 のツイッターでの投稿とブログでの投稿を基に分析された性格では、利用している SNS は全く異なるが、概ね性格グラフの形が似ているように見える。それに対し、被験者 1 と被験者 2 では、性格グラフの形が全く異なっているように見える。また、被験者 1 はツイッター、ブログどちらのグラフにも共通して Openness において高い値を示しており、定性的に判断しても、確かに知的好奇心が強い人物である。さらに、被験者 2 については Agreeableness において高い値を示しており、こちらも定性的に判断しても、確かに協調性の高い人物である。

これらの結果から、Personality Insights では、同一人物の異なる文書から性格ベクトル化を行っても、同じような結果を出力し、定性的に判断しても、概ね正しい性格を出力しているように見え、本研究への適用可能性は十分にあると考えられる。

3. 属性抽出

本章では、従来研究における属性抽出との相違点を述べながら、本研究における属性抽出のタスクを明確化する。なお、手法の詳細については 4 章にて述べる。

3.1 従来の属性抽出手法

SNS からの属性抽出において、従来研究では、近藤ら [3] の位置情報付きツイートを用いたユーザ属性推定と地域クラスタリングに関する研究や、上里ら [4] のメンション情報を利用した Twitter ユーザプロフィール推定における単語重要度算出手法に関する研究などが存在する。前者の研究では SNS からユーザの性別や年代を抽出しており、後者の研究では SNS からユーザの趣味・趣向、肩書などの属性が抽出できている。

3.2 属性抽出のタスク

一方、本研究の最終目標としては、うつ病患者の性格をベクトル化し分析することで、症状の緩和や改善の手助けとなるような属性を提案することを想定している。また、うつ病の患者の性格ベクトルにも個人差があると想定し、出来る限りその人の性格ベクトルに合った属性を付与できるようにするために属性を網羅的に取得することがタスクの1つであり、男女や年代のみの属性抽出では本タスクには対応できない。

さらに、網羅的に人間の様々な要素を属性として抽出しようとした場合、個人に依ってその位置付けが異なる属性が生じる。例えば、同じ“親”という属性を持つ人間がいたとしても、その“親”が自身に子供がいることを示すのか、もしくは自身に親がいることを示しているのかによって、“親”という属性の意味が全く変わってしまうのである。これでは属性が性格に及ぼす作用の存在解明を行う際に、位置付けの異なる2つの属性を1つの属性として扱ってしまうという問題が生じる可能性がある。そのため、属性抽出時に、ある属性が各ユーザに対してどういった位置付けにあるのかという情報が無い従来研究の手法では、本タスクに対しては不十分である。

そのため本研究では、同じ属性語にも複数の位置付けがあることを考慮し、属性定義を大きく4つのカテゴリ c_k に分け、それぞれを趣味・趣向（以降Hobbyまたは c_1 と呼称）、所属（以降Belongまたは c_2 と呼称）、所有（以降Ownershipまたは c_3 と呼称）、習慣（以降Customまたは c_4 と呼称）と定義した。

図3は、属性抽出の大まかな流れを示している。まず(1)にて、属性でありそうな単語をリストアップし、属性候補リストを生成する。次に(2)にて、生成した属性候補リスト内の各単語 w_j に対し、その単語が属性カテゴリ c_k に含まれるかどうかのラベル付けを行い、これをペア (w_j, c_k) とし、最後に(3)にて、ペア (w_j, c_k) がユーザ u_i の属性かどうかを判断する。

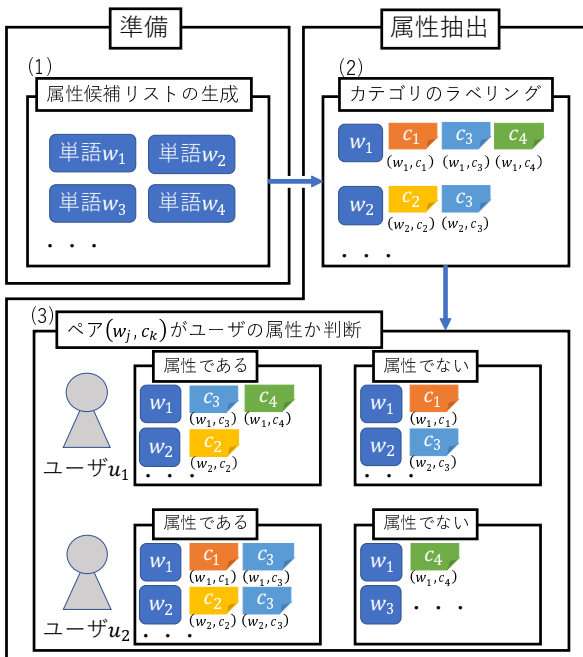


図3 属性抽出の概要

4. ラベリングされた属性の抽出方式

本章では、属性カテゴリ c_k にラベリングされた単語 w_j (ペア (w_j, c_k)) がユーザ u_i の属性であることを次式にて議論する。

$$\text{score}(u_i, w_j, c_k) = \begin{cases} 1 & \text{if 属性カテゴリ } c_k \text{ にラベリング} \\ & \text{された単語 } w_j \text{ (ペア } (w_j, c_k)) \\ & \text{はユーザ } u_i \text{ の属性である} \\ 0 & \text{otherwise.} \end{cases}$$

著者らは、 $\text{score}(u_i, w_j, c_k)$ が1を持つ時の条件の違いに基づいて、2種類に大別される手法(4.1節, 4.2節)を提案する。

なお、実験対象のSNS、使用した属性候補リスト、正解セットの作成等については、5.1節の実験環境にて記述する。

4.1 パターンマッチに基づく手法 (PTN)

例えば、あるユーザのSNSの中に「私の趣味はサッカーだ」という表現が含まれていた場合を考える。この場合我々は、単語「サッカー」には属性カテゴリ「趣味」がラベリングされるという点、さらに単語「サッカー」はそのユーザの属性であるという点の2点を同時に推定することができる。本節の手法はこの考えに基づき、ある単語 w_j が、属性カテゴリ c_k にラベリングできることが明示された表現で、ユーザ u_i のSNSの中に出現する時、属性カテゴリ c_k にラベリングされた単語 w_j (ペア (w_j, c_k)) はユーザ u_i の属性であると判別する。本稿では、ある単語 w_j が、属性カテゴリ c_k にラベリングできることが明示された表現について、表1のようにパターン化した。なお、この他にも各属性カテゴリに対していくつかのパターンが存在し、パターンごとにも主格や格助詞、文末表現などにバリエーションが存在する。

表1 属性カテゴリ c_k におけるレベル別のパターン例

c_k	Lv.3	Lv.2	Lv.1
趣味	私の趣味は xxx です	趣味は xxx です	趣味は xxx
所属	私は xxx 在住です	xxx 在住です	xxx 在住
所有	私は xxx を買いました	xxx を買いました	xxx を買う
習慣	私は毎日 xxx します	毎日 xxx します	毎日 xxx
		xxx : 一般名詞または固有名詞	

まず、表1のLv.3では、そのパターンがユーザの u_i のSNSの中に出現した時、単語 w_j に属性カテゴリ c_k がラベリングされる点と、単語 w_j はユーザ u_i の属性であるという点の2点の双方が十分に認められるようなパターンを選定した。

次に、表1のLv.2では、単語 w_j に属性カテゴリ c_k がラベリングされる点は認められるが、単語 w_j がユーザ u_i の属性であることは断定できないようなパターンを選定した。例えば「趣味はサッカーです」の場合、文の主体が不明であるため、単語「サッカー」はそのユーザの属性であることを断定できない。

最後に、表1のLv.1では、単語 w_j に属性カテゴリ c_k がラベリングされる点と、単語 w_j はユーザ u_i の属性であるという点の双方とも断定できないようなパターンを選定した。例えば「趣味はサッカー」の場合、Lv.2と同様のことが言える他、

「だ」「です」等の断定表現も不明であるため、単語「サッカー」に属性カテゴリ「趣味」がラベリングされる点も断定できない。

従って本稿では、システムが $\text{score}(u_i, w_j, c_k)$ に 1 を与える条件を次式の通りモデル化した。但し $\text{hit}_i(u_i, w_j, c_k)$ は、単語 w_j が、属性カテゴリ c_k にラベリングできるようなレベル $\text{Lv.}l$ のパターンが、ユーザ u_i の SNS 内で出現した回数を表す。

$$\text{score}(u_i, w_j, c_k) = \begin{cases} 1 & \text{if } \text{hit}_i(u_i, w_j, c_k) \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

4.2 単語間の関連度及び単語の出現回数に基づく手法

前節の手法は、単語 w_j に属性カテゴリ c_k がラベリングされるという点と、単語 w_j はユーザ u_i の属性であるという点の 2 点を同時に推定できる言語的表現を検討した上で、その表現がユーザ u_i の SNS に含まれているか否かに着眼したものであった。これに対し本節の手法は、単語 w_j はユーザ u_i の属性であるか否かという点と、単語 w_j に属性カテゴリ c_k がラベリングされるか否かという点についてそれぞれ別々に評価し、双方とも条件を満たす場合、属性カテゴリ c_k にラベリングされた単語 w_j (ペア (w_j, c_k)) はユーザ u_i の属性であると判別する。本稿では、前者について、単語 w_j がユーザ u_i の SNS に出現した回数 $\text{wc}_{u_i}(w_j)$ が閾値 cth_c 以上の場合、条件を満たしたと判断する。また後者については、単語 w_j と属性カテゴリ c_k との関連度 $\text{related}(w_j, c_k)$ が閾値 rth_c 以上の場合、条件を満たしたと判断する。なお、閾値は属性カテゴリ c_k に依存する。

$$\text{score}(u_i, w_j, c_k) = \begin{cases} 1 & \text{if } \text{wc}_{u_i}(w_j) \geq \text{cth}_c \\ & \text{and } \text{related}(w_j, c_k) \geq \text{rth}_c \\ 0 & \text{otherwise.} \end{cases}$$

ここで著者らは、関連度 $\text{related}(w_j, c_k)$ の算出方式の違いに基づいて、2 種類に大別される手法を提案する。例えば、単語「サッカー」と属性カテゴリ「趣味」の関連度を算出するケースを考える。1 つ目の手法は、「サッカー」と「趣味」が意味的に類似しているかという指標から算出するアプローチである。もう 1 つの手法は「趣味はサッカーです」のような言語的表現に違和感があるか否かという指標から算出するアプローチである。以降それぞれの詳細を 4.2.1 項と 4.2.2 項で述べていく。

4.2.1 単語間の意味的な類似度に基づく手法 (W2V)

本手法では、Word2Vec にて、単語 w_j と属性カテゴリ c_k の関連度 $\text{related}(w_j, c_k)$ を算出する。但し、属性カテゴリ c_k の連想語 $c'_{k,m}$ と単語 w_j との類似度 $\text{sim}(w_j, c'_{k,m})$ に着目している。

例えば、姉を持つあるユーザの属性として、属性カテゴリ「所有」がラベリングされた単語「姉」を抽出したい場合に、単語間の類似度を算出する手法の多くは、それらの共起性に着目しているため、単語「姉」と属性カテゴリ「所有」の意味的な類似度は必ずしも高くなるとは限らない。しかしながら、「所有」の連想語 (例えば「家族」や「持つ」など) と「姉」との類似度であれば、高くなることも十分に予想できる。

そこで本稿では、単語 w_j (但し、単語 w_j は Word2Vec のモデルに存在していることとする) と属性カテゴリ c_k との関連度

表 2 各属性カテゴリ c_k に対する連想語 $c'_{k,m}$ の集合 $A(c_k)$

属性カテゴリ c_k	連想語 $c'_{k,m}$ の集合 $A(c_k)$
趣味	スポーツ グルメ アニメ マンガ ゲーム 音楽
所属	集団 学生 職業 大人 子供 スポット 年代
所有	もの 家族 持つ 使う 買う
習慣	毎日 いつも する

$\text{related}(w_j, c_k)$ を 3 つのステップから算出する。まずステップ 1 では、属性カテゴリ c_k の連想語 $c'_{k,m}$ の集合 $A(c_k)$ をヒューリスティックに定める (表 2)。次にステップ 2 では、属性カテゴリ c_k の各連想語 $c'_{k,m}$ と単語 w_j との類似度 $\text{sim}(w_j, c'_{k,m})$ を算出する。最後にステップ 3 では、ステップ 2 で算出した類似度 $\text{sim}(w_j, c'_{k,m})$ の中で、最も高かったものを、単語 w_j と属性カテゴリ c_k の関連度 $\text{related}(w_j, c_k)$ として与える。

$$\text{related}(w_j, c_k) = \begin{cases} \max \{ \text{sim}(w_j, c'_{k,m}) \mid c'_{k,m} \in A(c_k) \} & \text{if the model of Word2Vec has } w_j \\ -\infty & \text{otherwise.} \end{cases}$$

4.2.2 単語とカテゴリの共起表現に基づく手法 (SoT)

例えば「趣味はサッカーだ」の表現は自然であり、単語「サッカー」に属性カテゴリ「趣味」がラベリングされることに違和感はない。すなわち、単語「サッカー」と属性カテゴリ「趣味」の共起表現「趣味はサッカーだ」には一般性があると言える。一方「趣味は姉だ」の表現は不自然であり、単語「姉」に属性カテゴリ「趣味」がラベリングされることに違和感がある。本項の手法はこの考えに基づき、単語 w_j と属性カテゴリ c_k の共起表現に対する一般性 (どれだけ自然な表現か) を利用し、単語 w_j と属性カテゴリ c_k の関連度 $\text{related}(w_j, c_k)$ を算出する。本稿では、単語 w_j と属性カテゴリ c_k の共起表現に対する一般性を、その表現が含まれる一般ユーザのツイート数に基づき算出する手法を定義する。

表 3 各属性カテゴリにおける単語 xxx との共起表現パターン

属性カテゴリ	共起表現パターン	
趣味	趣味は xxx です	私は xxx が好きです
所属	私は xxx 所属です	私は xxx 在住です
所有	私には xxx があります	私は xxx を持っています
習慣	いつも xxx します	毎日 xxx します

まず、各属性カテゴリ c_k に対して、2 つの共起表現パターンを用意する (表 3)。次に、全て (8 種類) の共起表現パターンに対して、ある単語 w_j を xxx に当てはめ、その表現が含まれるツイート件数を取得する。但し、取得には TwitterAPI の完全一致検索を用い、1 つの共起表現パターンの最大取得件数は 15 件とする。最終的に、単語 w_j と属性カテゴリ c_k における、1 つ目の共起表現の取得件数と 2 つ目の共起表現の取得件数の合計値 $\text{pc}_{1+2}(w_j, c_k)$ を 30 で除算した値が、単語 w_j と属性カテゴリ c_k の関連度 $\text{related}(w_j, c_k)$ として与えられる。

$$\text{related}(w_j, c_k) = \frac{\text{pc}_{1+2}(w_j, c_k)}{15 \cdot 2}$$

表 4 属性抽出 3 手法における属性カテゴリごとの最高 F 値
(ヒット数/システム出力数/正解セット数)

	Hob.	Bel.	Own.	Cus.	All
PTN	0.014 1/3/141	0.031 1/3/62	0.192 5/15/37	0.000 0/2/64	0.043 7/23/304
W2V	0.454 77/198/141	0.413 19/30/62	0.274 10/36/37	0.276 20/81/64	0.388 126/345/304
SoT	0.390 47/100/141	0.146 6/20/62	0.172 5/21/37	0.275 15/45/64	0.300 73/186/304

5. 評価実験

本章ではまず、実験環境について述べる。その上で、全手法について属性カテゴリごとの最高 F 値を示し、これについて考察する。その後、各 3 手法についてそれぞれパラメータを変化させた際の F 値に関する比較実験を行い、これらを考察する。

5.1 実験環境

本節では、属性候補リストの作成、正解セットの作成、各手法のパラメータ設定等の実験環境について述べる。

5.1.1 属性候補リスト

属性候補リストには、ツイッターから無作為に取得した 1000 人のプロフィール (以降 Bio とする) を mecab-ipadic-NEologd 辞書 [5] を用いた MeCab で形態素解析した結果の名詞 (固有名詞の人名は除く) の中から、さらに 2 人以上のユーザが Bio 内で使用している単語のみリスト化したものを用いる。

5.1.2 正解セット

本稿の実験では、ツイッターを利用している被験者 5 人に対し、被験者のユーザ投稿と Bio 内に含まれる単語をリスト化したものと属性候補リストの両方に含まれる単語のみをリスト化し、これを被験者に提示し、その中で被験者が自らの属性であると思う単語を全て選択させた。また、それらの属性がその被験者にとって、どの属性カテゴリにラベル付けされるかも重複可で設定させ、これらをその被験者の正解セットとした。

5.1.3 各手法のパラメータ

手法 PTN に対しては、対象とする文書をユーザ u_i の投稿に固定し、Lv.1 のみをパラメータとして与えた。

手法 W2V, SoT に対しては、対象とする文書をユーザ u_i のツイートと Bio, この 2 つを 1 つの文書として結合した 3 種類のドキュメント及び、単語カウントに関する閾値 cth_c 、関連度に関する閾値 rth_c をパラメータとして与え、手法 W2V の Word2Vec のモデルには Wikipedia モデル [6] を用いた。

5.2 3 手法への F 値の評価

本節では、各手法に対して、それぞれのパラメータを変化させた際の、各手法における最高 F 値を属性カテゴリごとに示した表 4 に関し考察を行う。なお、列 All は各属性カテゴリにおいて最高 F 値を得た際のヒット数、システム出力数、正解セット数を合算し、再計算した全属性カテゴリでの F 値である。

まず、表 4 より、All の F 値では、手法 PTN よりも、手法 W2V と手法 SoT の値が大きく上回った。このことから、手法

表 5 手法 PTN における各 Level と属性カテゴリごとの F 値
(ヒット数/システム出力数/正解セット数)

	Hob.	Bel.	Own.	Cus.	All
Lv.1	0.014 1/3/141	0.031 1/3/62	0.192 5/15/37	0.000 0/2/64	0.043 7/23/304
Lv.2	0.000 0/0/141	0.000 0/0/62	0.000 0/1/37	0.000 0/0/64	0.000 0/1/304
Lv.3	0.000 0/0/141	0.000 0/0/62	0.000 0/0/37	0.000 0/0/64	0.000 0/0/304

W2V と手法 SoT に共通するアプローチである、単語 w_j と属性カテゴリ c_k との関連度を測り属性カテゴリごとにその閾値を設定したり、単語 w_j のカウント数の閾値を属性カテゴリごとに設定することには一定効果があると考えられる。また、全手法に共通して Ownership と Custom の F 値が低く、これは、属性候補リストに問題がある可能性が大きい。理由としては、属性候補リストはユーザの Bio 群を元に生成しているが、ユーザが Bio に記述する単語として、趣味や所属は一般的であるが、習慣や所有については、前述した 2 つと比べるとあまり多くない印象があるからである。以上より、今後全体の精度を上げていくためには、属性候補リスト生成に用いるドキュメントの再考や、属性候補リストを用いない手法の提案等も必要である。

5.3 手法 PTN に関する F 値のパラメータ依存性

本節では、手法 PTN に対しパラメータを変化させた際の、最高 F 値を属性カテゴリごとに示した表 5 に関し考察を行う。

表 5 より、まず、Lv.2, Lv.3 では高い適合率を示すことを予想していたが、本実験での被験者 5 名の中ではこれらのパターンが殆ど出現せず、正解に対するヒット数も全属性カテゴリにおいて 0 回であり、Lv.2, Lv.3 における適合率の高さは認められなかった。これは、ツイッターで書かれる文章が口語的であったり、文章として成り立っていないことも多くあり、厳密なパターンの適用が難しいことが理由として考えられる。

また、Lv.1 に関しては、さらに多くの出力数を予想していたが、他の 2 手法に比べ出力数が少なかった。このことから、今後もパターンを利用するためには、ツイッター特有の言語表現からパターンを再考する必要があると考えられる。

5.4 手法 W2V に関する F 値のパラメータ依存性

本節では、手法 W2V で各属性カテゴリにおいて最高の F 値を得た際のパラメータを使用し、単語カウントに関する閾値 cth_c のみを変化させた場合の F 値の変化を示した図 4 と、関連度に関する閾値 rth_c のみを変化させた場合の F 値の変化を示した図 5 の 2 つについて考察を行う。なお、全カテゴリにおいて最高 F 値を示した際のドキュメントはツイートであった。

まず、図 4 より、Custom カテゴリ以外の属性カテゴリについては、単語カウントに関する閾値 cth_c が高くなるにつれてその F 値が下がっていくが、Custom カテゴリのみ単語カウントに関する閾値 cth_c が 10 付近の際に F 値のピークを迎えている。これより、Custom (習慣) の特性上、その判定にはやはり単語カウントに関する閾値が必要であることが分かった。

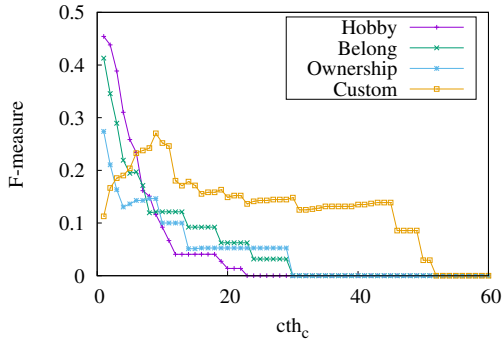


図4 手法 W2V における閾値 cth_c の変化に伴う F 値の変化

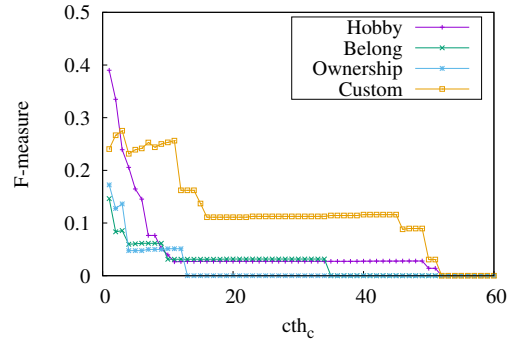


図6 手法 SoT における閾値 cth_c の変化に伴う F 値の変化

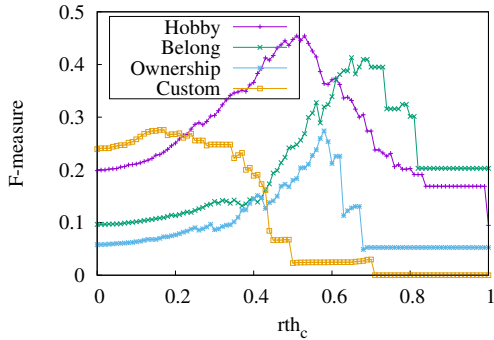


図5 手法 W2V における閾値 rth_c の変化に伴う F 値の変化

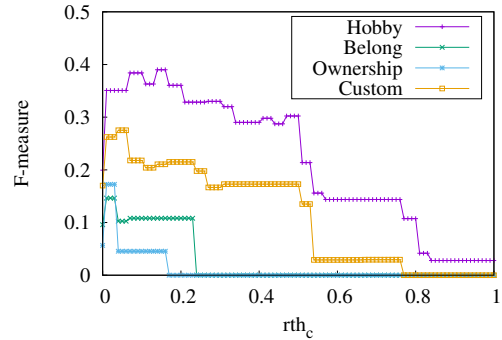


図7 手法 SoT における閾値 rth_c の変化に伴う F 値の変化

また、図5より、Custom カテゴリと単語 w_j との関連度は0.2以下の際に F 値がピークを迎えており、Custom カテゴリに関しては、他の属性カテゴリとは違い、属性カテゴリとの関連度よりもその出現回数が重要であることが分かった。その他の属性カテゴリに対しては、図5より適切な関連度に関する閾値 rth_c が存在し得ることが窺える。

5.5 手法 SoT に関する F 値のパラメータ依存性

本節では、手法 SoT で各属性カテゴリにおいて最高の F 値を得た際のパラメータを使用し、単語カウントに関する閾値 cth_c のみを変化させた場合の F 値の変化を示した図6と、関連度に関する閾値 rth_c のみを変化させた場合の F 値の変化を示した図7の2つについて考察を行う。なお、全カテゴリにおいて最高 F 値を示した際のドキュメントはツイートであった。

まず、図6より、Custom カテゴリに関しては、 w_j の単語カウントの閾値 cth_c が10を過ぎた辺りから、大きく F 値が下がっており、5.4節での考察と同様に、やはり単語カウントの閾値が重要であることが分かる。

また、図7より、Belong カテゴリと Ownership カテゴリの F 値は、他の2つの属性カテゴリよりも常に大きく下回っていることが分かり、これはこの2つの属性カテゴリにおける TwitterAPI でのパターン検索結果数が少ないことを意味し、手法 SoT の適用が困難であることが分かった。

6. まとめと今後の課題

本稿では、属性が性格に及ぼす作用の存在解明のための、Personality Insights を用いた性格ベクトル化手法と、属性抽出のためのパターンベース手法、Word2Vec の類似度を用いた手法、ツイッター検索の結果を用いた手法の4手法を提案した。

さらに本稿から、カテゴリ付けした属性の抽出には、Word2Vec の類似度を用いた手法で最も高い精度を示すことが分かった。また、属性カテゴリごとに単語の関連度や、出現回数の閾値に対する F 値の変化に特徴があることが分かった。

今後は本稿では行えなかった、ツイッターデータをコーパスとした際の Word2Vec の類似度を用いた手法を再実験すると共に、Word2Vec の類似度を用いた手法とツイッター検索の結果を用いた手法の2種類の関連度を組み合わせ、2種類の提案手法のハイブリッドを実験し、精度向上を図り、属性が性格に及ぼす作用の存在解明を目指す。

文 献

- [1] IBM Watson Personality Insights - Japan, <https://www.ibm.com/watson/jp-ja/developercloud/personality-insights.html> (2019).
- [2] 川本 哲也, 小塩 真司, 阿部 晋吾, 坪田 祐基, 平島 太郎, 伊藤 大幸, 谷 伊織, “ビッグ・ファイブ・パーソナリティ特性の年齢差と性差: 大規模横断調査による検討,” 発達心理学研究, Vol.26, No.2, pp.107-122 (2015).
- [3] 近藤 聖也, 吉田 孝志, 和泉 潔, 山田 健太, “位置情報付きツイートを用いたユーザ属性推定と地域クラスタリング,” The 30th Annual Conference of the Japanese Society for Artificial Intelligence, 1H2-1 (2016).
- [4] 上里 和也, 田中 正浩, 浅井 洋樹, 山名 早人, “メンション情報を利用した Twitter ユーザプロフィール推定における単語重要度算出手法の考察,” 情報処理学会研究報告, Vol.2014-DBS-159, No.22, Vol.2014-IFAT-115, No.22, pp.1-6 (2014).
- [5] Toshinori Sato, GitHub - neologd/mecab-ipadic-neologd: Neologism dictionary based on the language resources on the Web for mecab-ipadic, <https://github.com/neologd/mecab-ipadic-neologd> (2019).
- [6] カメリオ開発者ブログ, word2vec の学習済み日本語モデルを公開します, <https://aial.shiroyagi.co.jp/2017/02/japanese-word2vec-model-builder/> (2019).