

Twitter を用いた群衆のパーソナリティのクラスタ分析による県民性抽出

若宮 翔子[†] 李 龍[‡] 河合 由起子[†] 角谷 和俊[¶]

[†] 京都産業大学 〒603-8555 京都市北区上賀茂本山

[‡] Korea Institute of Science and Technology Information (KISTI), KOREA

[¶] 兵庫県立大学 〒670-0092 兵庫県姫路市新在家本町 1-1-12

E-mail: [†] s.wakamiya@gmail.com, kawai@cc.kyoto-su.ac.jp, [‡] lee.ryong@gmail.com,

[¶] sumiya@shse.u-hyogo.ac.jp

あらまし 地域の特徴を理解することは、引っ越し、マーケティングや広告のような地理的な意思決定を行う様々な場面において重要である。従来、社会調査のデータに基づく地域特徴の分析・抽出が行われている。しかしながら、社会調査に基づく地域特徴は、頻繁に実施することが困難であるためリアルタイム性が低いという問題や、大規模な地域の人々を対象とした調査が難しいという地理的な拡張性に関する問題がある。これに対し、近年普及している位置ベース SNS を活用することで、大規模な群衆の経験をリアルタイムかつ客観的に観察することができるようになり、従来の社会調査に基づき抽出される地域特徴を補完することが可能である。本稿では、地域特徴の一つとして地域の人々に共通する気質や性格のパロメータである県民性に着目し、これを抽出するためのプラットフォームを提案する。具体的には、Twitter から収集した群衆のライフログのメタデータに基づき各ユーザの行動を抽出する。そして、行動に基づくパーソナリティを分析することによりユーザのタイプを判定し、ユーザごとに社会的遺伝子リストを抽出する。次に、ユーザを実空間にマッピングし、ユーザ間の物理的な近接性とタイプの類似に基づく地理的クラスタリングを行う。最後に、地域クラスタごとに特徴的なユーザタイプをそのクラスタの県民性として特定する。これにより、行政区画に基づく都道府県だけでなく、より小さい地理的領域あるいはより大きい地理的領域において共通の人々の行動パターンを発見することが可能になる。さらに、Twitter を用いてユーザの行動を客観的に観察することで、従来の社会調査とは異なる観点に基づく地域特徴を把握することが可能になると期待される。

キーワード 県民性, パーソナリティ, 地理的クラスタリング, Twitter, 位置ベース SNS

1. はじめに

地域の特徴を理解することは、引っ越し、マーケティングや広告といった私生活やビジネスにおける地理的な意思決定を行う様々な場面において重要である。従来、地域特徴の抽出・分析を行うために、質問票を用いたアンケート、インタビューや観察といった社会調査によって得られたデータが用いられている。例えば、Rentfrow ら [5] はアメリカに住む 150 万人に対してパーソナリティテストを実施し、その結果を集約してアメリカの州ごとに共通する人々の特徴（県民性）を分析している。しかしながら、このような社会調査は、頻繁に実施することが困難であるため常にリアルタイム性の高い結果を得ることが難しいという問題や、大規模な地域における大量の人々を対象とした調査が難しいという拡張性に関する問題がある。また、ユーザの主観に依存することが多く、実際の地理的な意思決定の場面において客観性に欠ける結果となる可能性がある。しかしながら、従来の社会調査の観察によるアプローチでは、大量の人々を対象とすることが難しいだけでなく、客観的に観察することができる経験が限定されるという問題がある。

これに対し、位置ベース SNS における群衆のライフログを活用することで、従来の社会調査に基づき抽出される地域特徴を補完することが可能である。実際に、Twitter のような位置ベース SNS は、人々の実世界における日常的な活動をリアルタイムに発信・共有することができる場として普及しており、リアルタイム性が高く、広域にわたる大量の人々の様々な経験を客観的に観察することが可能である。このような位置ベース SNS の利点に着目し、我々はこれまでに、社会的地域イベントの検出 [1]、地域特徴の抽出 [2] や社会的地域認知の測定 [7] といった群衆ベースの地域分析に関する研究を行ってきた。

本稿では、Twitter を用いた地域分析の一つとして、各ユーザの日常的なパターンを考慮した県民性の抽出に着目する。一般的に、県民性とは、集団の大多数に共通する性質であるモータル・パーソナリティ（最頻性格）の一部であり、都道府県を単位とした人々のモータル・パーソナリティとして定義される [8]。しかしながら、都道府県は行政のために定められた地理的単位であり、都道府県よりもさらに小さい地域や、複数の都道府県にまたがる地域に共通する性質も存在す

と考えられる。本稿では、地理的に近接している人々に共通する社会的遺伝子を県民性と定義し、これを抽出するためのプラットフォームを提案する。具体的には、Twitter から収集した群衆のライフログのメタデータに基づき各ユーザの行動を抽出する。次に、行動に基づきユーザのパーソナリティを測定して、分析してユーザのタイプを判定し、各ユーザの社会的遺伝子リストを抽出する。そして、ユーザを実空間にマッピングし、ユーザ間の物理的な近接性と社会的遺伝子リストの類似に基づく地理的クラスタリングを行う。最後に、地域クラスタにおけるユーザに共通の社会的遺伝子をその地域クラスタの県民性として特定する。

本論文の構成は以下の通りである。まず、2 章で本研究のモデルを紹介し、関連研究について述べる。次に、3 章では提案手法の具体的な手順を説明する。4 章で本論文のまとめと今後の計画と課題について述べる。

2. 本研究のアプローチ

本章では、Twitter を用いた県民性抽出のための研究モデルを紹介し (2.1 節)、関連研究を要約して、提案手法との差異を明らかにする (2.2 節)。

2.1 研究モデル

本研究では、Twitter を地理空間・時間・人々の経験からなる多次元のデータベースとみなしてモデリングし、地理社会分析のために活用することを目的とする。本稿では、各ユーザの行動やメンタルのパターンを考慮した県民性を抽出するためのプラットフォームを提案する。図 1 に提案する研究モデルを示す。直感的に、我々は、一般的なユーザの行動やメンタル状態のパターンを社会的遺伝子とみなして抽出し、各ユーザにとって優性な遺伝子をパーソナリティに基づき判定する。図 1 では、ユーザ A から F の 6 名のツイートから抽出した各ユーザの社会的遺伝子をレーダーチャートにより表現している。各項目はユーザのパーソナリティに基づき判定されるユーザタイプとする。この社会的遺伝子が類似し、かつ地理的に近接しているユーザをクラスタリングし、共通する社会的遺伝子を抽出することで、その地域クラスタの県民性を判定する。

本稿で取り組むべき技術的なポイントは以下の通りである。

- 1) 大量のストリーミングライフログの段階的処理によるユーザの社会的遺伝子の抽出
- 2) ユーザの行動拠点分析によるユーザと地域の関連性判定
- 3) 地理的クラスタリングによる県民性抽出

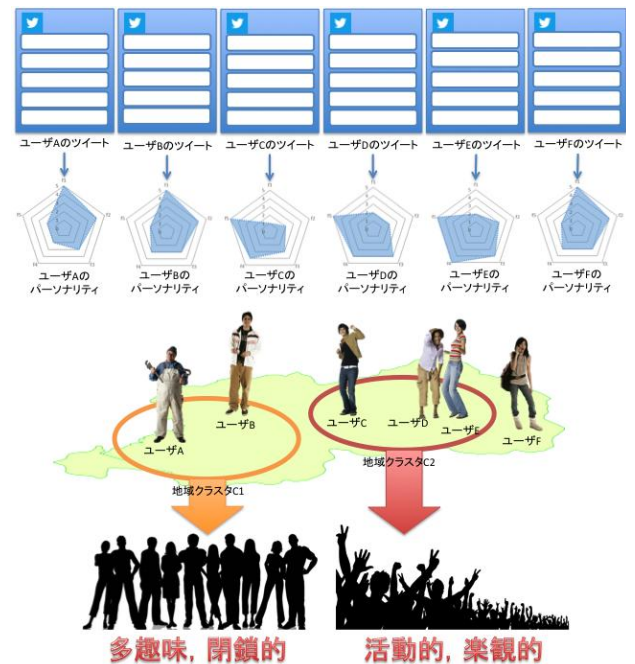


図 1. Twitter を通したユーザのパーソナリティの地理的クラスタ分析による県民性抽出

2.2 関連研究

SNS に蓄積されている群衆のライフログを活用して、ユーザを分類する研究が盛んに行われている [4]。また、分類したユーザを対象として、意見調査や評判調査をする研究も行われている。我々の過去の研究 [6] では、Twitter を用いたテレビ視聴率調査を実現するために、ツイートとテレビ番組の関連性を内容、時間、場所の観点で判定することにより、各テレビ番組の視聴者を特定する手法を開発した。この手法では、単に番組の視聴者を集計することで番組の特徴付けを行ったのに対し、本研究では、ユーザの日常的な行動に基づき判定されるユーザタイプに基づく地理的なクラスタリングを行うことで県民性を抽出している。

さらに、我々は位置ベース SNS を活用し、地理社会分析について検討してきた。これらの研究では、Twitter を地理空間・時間・人々の経験からなる多次元データベースとしてモデリングしており、今回の県民性抽出についても共通のモデリングを行っている。地域特徴抽出に関する手法 [2] では、あらかじめ設定した地域クラスタ内に存在する群衆の振る舞いを観察し、その時間的なパターンを分析した。社会的地理認知の測定に基づく近接地域検索 [7] では、あらかじめ設定した地域クラスタ間の群衆移動を集約することにより、群衆ベースの近接性を測定した。これに対し、提案手法では、ユーザごとにパーソナリティを抽出し、地理的クラスタリングを行うことにより、共通のパーソナリティのパターンを持つ地域クラスタを検出していると

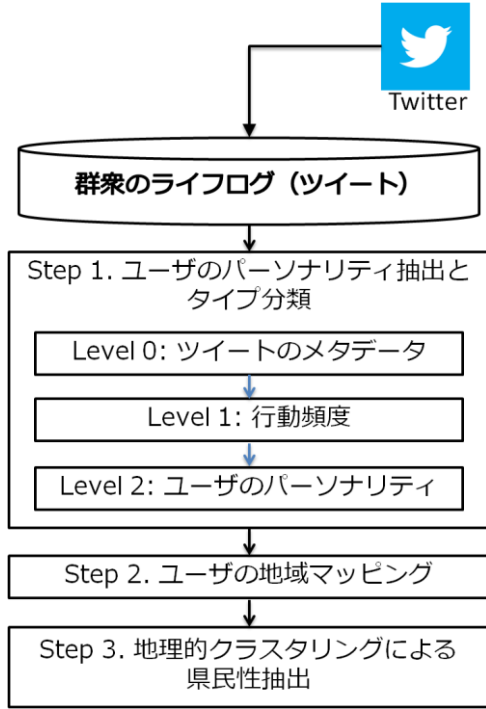


図 2. ツイートを用いた県民性抽出プロセス

いう点で、アプローチが大きく異なる。さらに、本研究では、Twitter を用いて県民性を特定することを目的としているという点についても異なる。

さらに、人々の感情を中心とした地域の性格を求めるために、Mislove [3] らは人々の感情の時間的な変化パターンを視覚的に表現した。具体的には、各地域のツイートを用いて感情を抽出し、各時間帯の感情の地理的分布に示すことで、アメリカのパルスを表現した。我々の研究では、Twitter を通してユーザのパーソナリティに基づくタイプを判定し、地理的に共通する人々のパターンを抽出している点で異なる。

3. 位置情報付きツイートを活用した県民性抽出手法

Twitter に投稿されている位置情報付きツイートをを用いて、地理的に共通する人々の特徴を県民性として求めるプロセスを具体的に説明する (図 2)。まず、Twitter から収集した群衆のライフログを用いてユーザごとにパーソナリティを抽出し、全ユーザのパーソナリティを分析することにより、ユーザタイプを判定し、ユーザの社会的遺伝子を抽出する (3.1 節)。次に、地理空間にユーザをマッピングし (3.2 節)、社会的遺伝子を地理的にクラスタリングすることにより、共通のユーザタイプを県民性として抽出する (3.3 節)。

3.1 ユーザのパーソナリティ抽出とタイプ判定

Twitter から収集したツイートから各ユーザのパーソナリティを抽出し、全ユーザのパーソナリティを集

約することにより、ユーザのタイプを判定する。本研究では、ユーザ $user_i$ のパーソナリティ $personality_i$ をユーザの行動 $behavior$ とメンタル状態 $mental$ からなるものとして定義する。大量のユーザがリアルタイムで発信するストリーミングデータの処理にはコストがかかるため、図 2 の Step1 のように段階的な情報抽出を行う。

Level-0: ツイートのメタデータを抽出

あるユーザ $user_i$ のツイート tw_j からメタデータを抽出する。このとき、ツイートに付与されているオリジナルのメタデータ ($time, loc (lat., long), msg$) に加え、メッセージをパースングし、Twitter における一般的な記述ルールにより簡単に抽出することができるデータ ($hashtag, url, reply, retweet$) を抽出する。各メタデータの定義は以下のとおりである。

- 時間: $time$
- 位置 (緯度・経度): $loc (lat., long)$
- テキストメッセージ: msg
- ハッシュタグ (Ex. "#hashtag"): $hashtag$
- URL (Ex. "https://twitter.com"): url
- 返信 (Ex. "@userID"): $reply$

例えば、ハッシュタグ、URL、返信は、「それぞれ '#'、'http://', '@' から始まる文字列」という Twitter における一般的な記述ルールに基づき、正規表現を用いて抽出する。

Level-1: ユーザの行動頻度を抽出

Level-0 で抽出したメタデータを組み合わせてユーザの経験 (行動やメンタル状態など) を抽出する。ユーザ単位での処理を行うため、ツイートのユーザ ID をもとにユーザごと $user_i (\in U)$ にメタデータ $Metadata_i$ を管理する。具体的には、新しいツイートを取得するたびに、重複のないユーザ ID リスト U とそのツイートのユーザ ID を比較し、リストに存在しない場合はそのユーザ ID を追加する。本稿では、以下に示すユーザの基本的な行動を抽出し、その頻度 $b_j (\in behavior)$ を求める。各行動と用いるメタデータとの関係を表 1 に示す。

- 移動: $move$

$$move = tw_j - tw_{j-1} \quad (1)$$

$$where \ tw_{j-1}.loc \neq tw_j.loc \text{ and } tw_{j-1}.time < tw_j.time \text{ and } tw_{j-1}, tw_j \in Tw_i$$

- 会話: $chat$
- 写真撮影: $photo$
- 動画視聴: $video$
- 外出: $checkin$
- テレビ視聴: tv
- 食事: $meal$

移動 $move$ は、式 (1) のように、時間的に連続するツ

表 1. メタデータを用いたユーザ行動の抽出

メタデータ 行動	時間 time	位置 loc	テキスト メッセージ msg	ハッシュ タグ hashtag	URL url	返信 reply
移動 move	●	●				
会話 chat						●
写真撮影 photo				●	●	
動画視聴 video				●	●	
外出 go-out					●	
テレビ tv			●			
食事 meal			●			

イトが異なる位置で発信された場合に抽出する．さらに，移動のメタデータとして，位置の差に基づく移動距離 *move_dist* や時間の差に基づく移動時間 *move_time* を算出することができる．会話 *chat* は他ユーザと関わる行動であるため，メタデータの返信 *reply* に基づき抽出する．写真撮影 *photo* や動画視聴 *video* は，メタデータの *url* や *hashtag* を分析することにより抽出する．具体的には，それぞれの行動のヒントとなる語の辞書に基づく正規表現を用いて判定する．写真撮影 *photo* や動画視聴 *video* を抽出するために定義した正規表現の例を以下に示す．

・ 写真撮影行動抽出のための正規表現の例

“http://twitpic.com/(¥¥S)+”

“http://www.flickr.com/(¥¥S)+”

“http://instagram/(¥¥S)+”

・ 動画視聴行動抽出のための正規表現の例

“http://youtube.com/(¥¥S)+”

“http://www.ustream.tv/(¥¥S)+”

“http://www.nicovideo.jp/watch/(¥¥S)+”

また，メタデータのテキストメッセージ *msg* を分析することによりテレビ視聴や食事といった具体的な行動を抽出することも可能である．

次に，ユーザの行動頻度をそれぞれ求める．行動頻度は，以下のように関数 *Count(.)* を用いて算出する．

- 移動頻度： *Count(move)*
- 会話頻度： *Count(chat)*
- 写真撮影頻度： *Count(photo)*
- 動画視聴頻度： *Count(video)*
- 外出頻度： *Count(checkin)*
- テレビ視聴頻度： *Count(tv)*
- 食事頻度： *Count(meal)*

Level-2：ユーザのパーソナリティを抽出

あるユーザ *user_i* のパーソナリティ *personality_i* を抽出

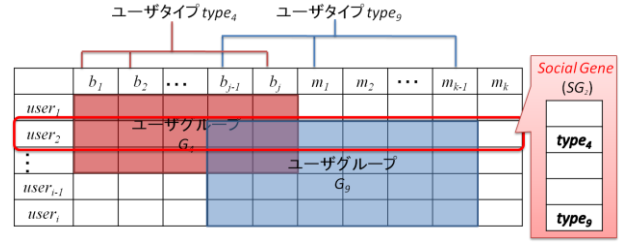


図 3. ユーザタイプに基づくユーザの社会的遺伝子リストの抽出

する．式 (2) のように行動頻度の集合 *behavior* とメンタル状態の集合 *mental* から構成されるベクトルとして表現する．

$$\text{personality}_i = \langle \text{behavior}, \text{mental} \rangle \quad (2)$$

$$(\text{behavior} \in B, \text{mental} \in M)$$

これにより，ユーザとパーソナリティからなる行列が生成される．この行列を分解することにより，ユーザのタイプを抽出する．具体的には，非負値行列因子分解アルゴリズム (NMF) や生物情報科学などで用いられている Biclustering などを用いて分析する．これにより抽出されるユーザタイプを社会的遺伝子と定義する．そして，ユーザごとに優勢なユーザタイプを求めることで，各ユーザの社会的遺伝子リストを抽出する．例えば，図 3 において，ユーザ・パーソナリティマトリクスを分析した結果， b_1, b_2, b_j からなるユーザタイプ *type₄* と b_{j-1}, m_1, m_{k-1} からなるユーザタイプ *type₉* を判定したとする．*user₂* のパーソナリティは，これらのユーザタイプに属していたとき，*type₄* と *type₉* を含むユーザタイプの集合をこのユーザの社会的遺伝子 *SG₂* として判定する．

$$SG_i = \langle \text{type}_1, \text{type}_2, \dots, \text{type}_{s-1}, \text{type}_s \rangle \quad (3)$$

$$\text{type}_s = \langle b_1, b_2, \dots, b_{j-1}, b_j, m_1, m_2, \dots, m_{k-1}, m_k \rangle (b_j \in B, m_k \in M)$$

3.2 ユーザの地域マッピング

3.1 節では，各ユーザのツイートを Level-0, Level-1, Level-2 と段階的に集約し，パーソナリティに基づくユーザの社会的遺伝子リストをそれぞれ抽出した．次に，ユーザがどの県民であるのか，すなわち，どの場所との関連性が高いのかを判定し，地理空間上にマッピングする．従来の社会調査では，居住している地域を回答してもらったり，地域を限定して調査したりすることによりユーザと地域を対応付けることができる．しかし，提案手法では，ツイートに基づくユーザの経験から活動の拠点としている場所を特定する必要がある．図 3 に Level-1 の前処理で抽出されるユーザの移動を 2 次元の地理空間上にマッピングした結果を示す．この図の縦軸は時間，色分けした線は個々のユーザの移動

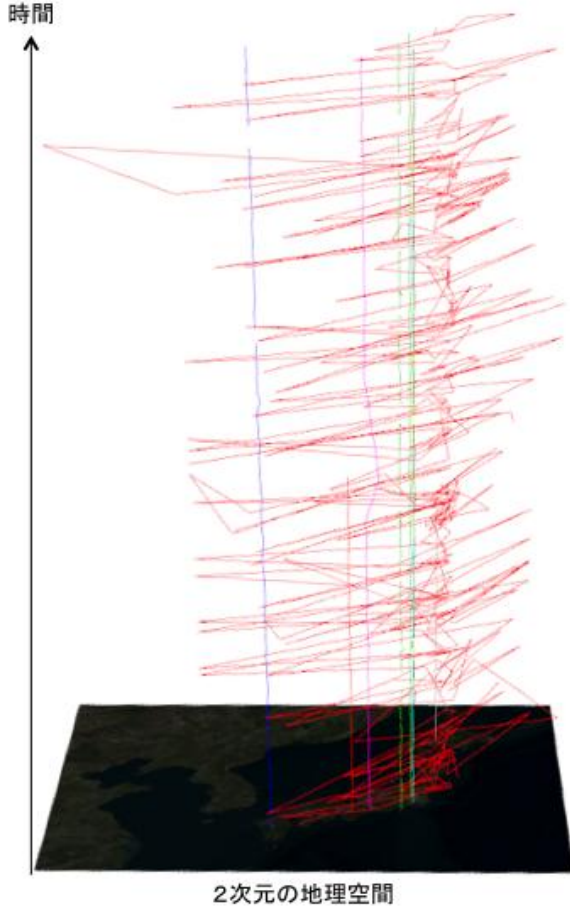


図 3. 多様なユーザの移動

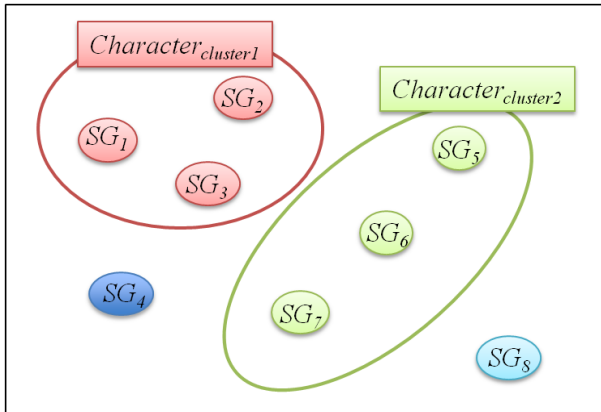


図 4. 地理的クラスタリングによる県民性抽出

を表している。青色と緑色のユーザの移動履歴は、最初にマッピングされた地点からほぼ垂直に上に向かっていていることから、大きく移動せず特定の場所でツイートを発信していたことが分かる。このようなユーザは、ツイートの地理的分布に基づき実空間にマッピングすることができる。しかし、これらの間に見られるピンク色のユーザの移動履歴は、一時的に異なる地域に移

動していることが分かる。この場合、一時的に訪れた地域を例外とみなすこともできる。しかし、このユーザが、普段職場でツイートを発信しており、たまたま自宅でツイートを発信したという可能性も考えられる。このような場合、単に頻度が多い場所にユーザをマッピングすることが正しいとは限らない。さらに極端な例として、赤色のユーザの場合、日本各地を頻繁に移動していることが分かる。このような場合、各地点における発信頻度を計算したとしても、その値が高くなる地域が複数出現すると考えられる。本稿では、このようなユーザと地域の関連性を判定することの難しさや解釈の多様性があることを承知した上で、単純にユーザが頻繁に訪れる地点にユーザをマッピングするものとする。

$$\begin{aligned} geo_i &= loc_j \\ \text{where } loc_j &\in Loc_i \text{ and } req(loc_j) \geq \alpha \end{aligned} \quad (4)$$

ここで、ユーザが訪れたことのある地点のうち、頻度が閾値 α 以上の場所をユーザ $user_i$ の位置情報 geo_i とする。

3.3 地理的クラスタリングによる県民性抽出

地理的に近接するユーザの社会的遺伝子リストの類似性に基づき、地域クラスタを設定し、共通の社会的遺伝子を県民性として抽出する。具体的には、Level-2で求めたユーザ $user_i$ の社会的遺伝子リスト SG_i を3.2節で求めた位置情報 geo_i をもとに地理空間上にマッピングし、すべてのユーザの社会的遺伝子を地理的にクラスタリングする。式(5)のように、地理的に近接しており、かつ社会的遺伝子リストの類似度が高いユーザを求めて地域クラスタを生成し、その地域クラスタのユーザに共通する社会的遺伝子を県民性として抽出する。

$$\begin{aligned} \text{Similarity}(user_i, user_j) \\ = w_{geo_dist} \times \frac{1}{geo_dist(geo_i, geo_j)} + w_{sim} \times sim(SG_i, SG_j) \quad (5) \\ \text{where } user_i, user_j \in Users \end{aligned}$$

4. 結論

本稿では、Twitterに投稿されている群衆のライフログを活用し、地理的に近接している人々に共通のユーザタイプを県民性として抽出するための手法を提案した。具体的には、大量のユーザのツイートを段階的に処理することで各ユーザのパーソナリティを抽出し、ユーザタイプに基づくユーザの社会的遺伝子を判定し、ツイート発信位置の地理的分布を考慮してユーザを地域にマッピングし、地域的クラスタリングでパーソナリティの特徴的なパターンを抽出する手法について述べた。これにより、行政区画に基づく都道府県だけで

なく、より小さい地理的領域あるいはより大きい地理的領域において共通の人々の行動パターンを発見することが可能になる。さらに、Twitter を用いてユーザの行動やメンタルを客観的に観察することで、従来の社会調査とは異なる観点に基づく地域特徴を把握することが可能になる。

今後の課題として、提案手法に基づき、Twitter から収集した位置情報付きツイートをを用いた実験を行う予定である。そして、抽出された県民性を評価するために、社会学的な観察や統計調査によるデータに基づく従来の調査結果と比較する。具体的な比較データとして、おもしろ県民性データブック¹や県民性の人間学(祖父江孝男 著)²などにまとめられている結果を利用することを計画している。特に、従来の手法では、あらかじめ決められた都道府県単位で人々の性格を調査することにより、県民性を調査している。これに対し、提案手法では、あらかじめ地域を指定するのではなく、パーソナリティが類似するユーザが存在する連続的な地域クラスターを求めている。そのため、従来の調査に基づく手法とは異なる県民性が抽出され、新しい観点での地域情報を抽出・提示することができると期待される。そのため、ある地域について、従来の調査結果で抽出された県民性と提案手法により抽出した県民性を対応付けることにより、提案手法によって補完できる観点について詳しく考察する予定である。また、時間帯を考慮したユーザの行動頻度やメンタル状態の測定を行う予定である。

謝辞

本研究の一部は、平成 24 年度特別研究員奨励費(24.9154) および総務省戦略的情報通信研究開発制度(SCOPE) によるものである。ここに記して謹んで感謝の意を表する。

参考文献

- [1] Ryong Lee, Shoko Wakamiya and Kazutoshi Sumiya, Discovery of Unusual Regional Social Activities using Geo-tagged Microblogs, *World Wide Web Journal Special Issue on Mobile Services on the Web*, Volume 14, Number 4, pp. 321-349, 2011
- [2] Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya, Urban Area Characterization based on Crowd Behavioral Lifelogs over Twitter, *Personal and Ubiquitous Computing*: Volume 17, Issue 4, pp. 605-620, 2013
- [3] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, J. Niels Rosenquist, Pulse of the Nation: U.S. Mood Throughout the Day inferred

from
<http://www.ccs.neu.edu/home/amislove/twittermood/>

- [4] Marco Pennacchiotti and Ana-Maria Popescu, Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter, In *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 430-438, 2011.
- [5] Rentfrow, P. J., Gosling, S. D., Stillwell, D. J., Kosinski, M., & Potter, J. (in press). Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, 2013.
- [6] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya, Crowd-powered TV Viewing Rates: Measuring Relevancy between Tweets and TV Programs, In *Proc. of the 2nd International Workshop on Social Networks and Social Media Mining on the Web (SNSMW 2011)*, Lecture Notes in Computer Science, 2011, Volume 6637, pp. 390-401, 2011.
- [7] Shoko Wakamiya, Ryong Lee and Kazutoshi Sumiya, Social-Urban Neighborhood Search based on Crowd Footprints Network, In *Proc. of the 5th International Conference on Social Informatics (SocInfo2013)*, pp. 429-442, Kyoto, Japan, November 25-27, 2013
- [8] 高橋正臣, モーダルパーソナリティとバーバルコミュニケーションの関連性について, 大分県立芸術文化短期大学 研究紀要 28, 1-21, 1990.

¹ おもしろ県民性データブック :
<http://diamond.jp/category/s-kenmin>

² 県民性の人間学(祖父江 孝男 著)

<http://www.chikumashobo.co.jp/product/9784480429933/>