3M4-4in

アイコン画像に注目したTwitter研究の提案

冨永　登夢　　土方　嘉徳　　西田　正吾
Tomu Tominaga　Yoshinori Hijikata　Shogo Nishida

大阪大学大学院　基礎工学研究科
Osaka University, Graduate School of Engineering Science

Recently, many researches focus on Twitter, which is a microblog service used by lots of people in the world. Twitter has some features such as real-time communications, limit on number of characters, usage of unique language on microblogs. Using these features, researchers have analyzed the links between users or the text users tweeted. Furthermore, they estimated users' preference or proposed the frameworks which can detect some event. As mentioned above, there are various researches concerning Twitter. However, there is no Twitter research focusing on the profile images of users. The profile images may imply users' sense of value, background, preference because users can select their favorite image or photo as the profile image. Thereupon, we divided empirically the profile images into 13 types, and researched the relationship between users' behavioral characters and the type of profile images.

1. はじめに

Twitter[*1] は，世界中の多くの人々に利用される代表的なマイクロブログサービスで，全世界で 1 億 4 千万のアクティブユーザが存在する [*2]．Twitter の持つ大規模なネットワークを対象に，ユーザプロフィールや投稿されたテキストの内容，ユーザ間のリンクなどから分析する研究が行われてきた．また，リアルタイム性，文字数制限によるテキストの簡略化，マイクロブログに見られる独特な言語の利用など，Twitter 特有の性質を利用した研究も多い．

我々がアイコン画像に注目する理由は，アイコン画像がユーザの内面的特徴を表す可能性を持つと考えるからである．アイコン画像とは，ユーザが自由に設定できるアカウントの画像や写真であり，いわば Twitter 上における自身の顔である．つまり，このアイコン画像を選択・設定する際には，個人の価値観や思想，嗜好などの影響を受けると考えられる．そこで我々はアイコン画像に注目し，ユーザの Twitter 上での行動とアイコン画像の関係を調査した．ここで得られた知見は，工学的にはユーザクラスタリングやコミュニティ抽出，社会学・心理学的にはネットワーク上におけるユーザの振る舞いに関する知識の発見などに貢献できると考えられる．ちなみに我々の知る限りでは，アイコン画像に注目した Twitter 研究は存在しない．

2. 従来の研究

マイクロブログサービスとは短いメッセージやテキストを投稿して他のユーザと交流するサービスのことである．その代表例として Twitter が挙げられる．他のユーザをフォローすることで，フォローした人の投稿内容を自身のタイムラインで閲覧できる．このようなマイクロブログの一つである Twitter は，以下のような有用な特徴と制約的な特徴を持つ．有用な特徴として挙げられるものは，リアルタイム性 [1] である．リアルタイム性とは，自分が意図したときに自分の持つ情報を迅速に発信できるという性質や，自分が情報を求めたときにその実時間情報を取得できるという性質のことである．例えば，天気情報を発信するユーザをフォローすることで天気予報をリアルタイムに取得したり，映画情報をつぶやくユーザから公開予定の映画の日程を確認したりできる．制約的な特徴として挙げられるものは，文字数制限によるテキストの簡略化，ミニブログ独特の言語の利用 [1] である．文字数制限によるテキストの簡略化とは，Twitter における投稿テキストは 140 字以内という字数制限があるため，投稿テキストの内容が顔文字やテキストスタンプ [*3] などで簡略化されることである．特に顔文字は，ユーザの感情を簡潔に表現するために使われることが多い．ミニブログ独特の言語の利用とは，音だけを当てはめて造語をしたり，極端な感情を表現するために長音化を行ったりすることである．例えば，'you' を 'u' と表記したり (省略)，'before' を 'b4' としたり (音声置換)，'goooooood' と書いて大げさな感情を表現する (長音化) ことが挙げられる．

これらの特徴と，1. 章で述べたように Twitter には大規模なネットワークが存在することを考慮して，Twitter の研究は進められている．このような分析を行った従来の研究を以下に紹介する．

2.1 実世界の動向分析

これは，Twitter 上でリアルタイムに発信される情報を利用して実世界の動向を分析する研究である．Bollen[3] らは，ツイートの感情を，POMS(Profile of Mood States) をもとにした 6 つの感情 (tension, depression, anger, vigor, fatigue, confusion) について分析した結果と，株式市場，原油市場，主要な出来事との関係を調査した．彼らは，世の中の様々なできことが感情に影響を与えていることを報告した．Asur と Huberman[4] は，ツイートが映画の興行収入を予測するのに利用できることを示した．Sakaki[5] らは，Twitter 上の実時間情報を調査し，イベントの発見を可能にした上で，地震の通知報告システムを提案した．

2.2 トピック同定

トピック同定とは，ユーザに対して，投稿内容に応じてタグの付与行ったり，関心の同定を行うすることである．投稿内容に応じたタグをユーザは，検索や推薦などに利用される他，それ自体をユーザの属性の一種として見なされることもある．

連絡先: 冨永登夢，大阪大学大学院基礎工学研究科システム創成専攻，大阪府豊中市待兼山町 1-3，06-6850-6382，tominaga@nishilab.sys.es.osaka-u.ac.jp

∗1　https://twitter.com/

∗2　平成 24 年度版 総務省情報通信白書第 1 部第 3 節

∗3　アスキーアート，もしくはアスキーアートを入力する機能

Pennacchiotti と Gurumurthy[6] は，LDA を用いてユーザをトピックの混合物として表現し，Twitter ネットワーク上から類似のユーザの推薦を行った．Zhao[7] らは，1 つのツイートは 1 つのトピックを持つ文書であるという仮説をもとに，LDA の拡張版である Twitter-LDA モデルを考案した．

2.3 信頼性評価

ユーザの投稿内容の信頼性を評価する研究について述べる．Castillo[8] らは，決定木学習を用いて Twitter の投稿内容が信頼できるかどうかを判定する分類器を構築した．Qazvinian[9] らは，ベイズ分類器を用いて，Twitter での信頼性に基づく噂検出を行った．

2.4 ユーザクラスタリング

ユーザクラスタリングとは，ある特徴を軸に類似するユーザ同士をあるクラスタにまとめることである．Bergsma[10] らは，ユーザの場所と名前 (ファーストネームとラストネーム) を素性として，コサイン類似度と K-means 法によりユーザをクラスタリングした．Pennacchiotti と Popescu[11] は，ユーザのプロフィール，利用履歴 (ツイートやリツイート，メンションなどの履歴)，投稿内容，そしてユーザ間のリンクを素性とし，GBDT(Gradient Boosted Decision Trees) によるクラスタリングの機械学習フレームワークを提案した．

3. アイコン画像の分類

アイコン画像の特徴とユーザ特性の関係を調査するために，アイコン画像を 13 種類の項目を定義し，経験的に分類した (以下ではこの項目のことを分類項目と呼ぶ)．動物，たまご，自画像，顔隠し，文字，ロゴ，オブジェ，オタク，本人一人，本人複数，景色，他人，キャラクタの 13 種類である．ここで，定義された 13 分類項目に対して，誰が分類しても同じ結果になるのか，Twitter ユーザのアイコン画像を網羅できるかという 2 点を検証するため，以下の 2 つの実験を行った．

3.1 分類の一致度評価

複数人による分類結果の一致度を測定する実験を行った．Twitter からランダムに選択したユーザ 300 人分のアイコン画像を，10 人の被験者に分類させた．この 10 人の分類結果を Siegel の一致係数で評価した．Siegel の一致係数とは，3 人以上の複数被験者の評価値がどの程度一致しているかを定量的に計測した値である．直感的には一般的なカッパ係数を 3 人以上の複数被験者のために拡張したものである．実験の結果，この係数の値は 0.70 となり，実質的に一致しているとみなされた．ちなみに，Siegel の一致係数の解釈は，0.00 ～ 0.40 で低い一致，0.41 ～ 0.60 で中程度の一致，0.61 ～ 0.80 で実質的に一致，0.81 ～ 1.00 でほぼ完全に一致となっている．

3.2 分類の網羅性評価

13 分類項目に対象ユーザが含まれない割合を調査した．まず，Twitter から Streaming API Sample エンドポイント *4 で日本語設定を行っているユーザから発信された 20833001 ツイートを取得した．これは 2013 年 9 月 18 日から 10 月 17 日の 1ヶ月間でクローリングを行って取得したものである．次に，それらに紐付けられた日本語ユーザからランダムに 1067 人選択した．これは社会標本調査における標本数の決定式 (1) で算出した人数である．

$$n = \frac{N}{(\frac{\epsilon}{\mu(a)})^2 \cdot \frac{N-1}{\rho(1-\rho)} + 1} \quad (1)$$

*4 https://dev.twitter.com/docs/api/1.1/get/statuses/sample

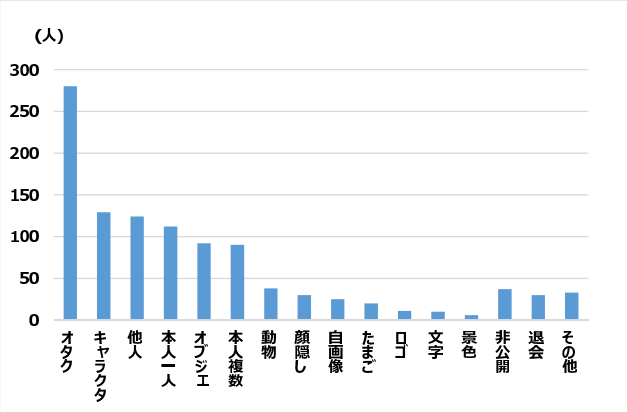

図 1: 網羅性評価

ここで，n は必要サンプル数，N は母数，$\mu(a)$ は信頼度 (100-a) %の時の正規分布の値，ϵ は精度，ρ は母比率である．一般的には，$\mu(a) = 1.96(a = 5)$，$\epsilon = 0.03$，$\rho = 0.5$ とされるため，本研究でもこれを用いた．N = 20833001 として必要サンプル数を求めると，n = 1067 となる．我々はデータセットとして，ツイートとそのツイートを行ったユーザの両者を持つが，各ユーザがどれだけツイートしたかは取得していない．本来，20833001 ツイートのデータセットの中には，同一ユーザのツイートが存在するため，実際のユーザ数は 20833001 より少ない．ただし，上記のパラメータは 1 人 1 ツイートのみ行ったと仮定したため，式 (1) で求めたサンプル数は，統計的には十分であると言える．従って，これにより選択された 1067 人分のユーザのアイコン画像の分類を行った．ここで，先述の 13 分類項目に加え「その他」項目を設けた．「その他」に含まれるユーザが少なければ，この分類による網羅性は高いと言える．また，非公開ユーザ *5 と退会ユーザ *6 は，調査対象外であるため分類の際に除いた．これらのユーザは，1ヶ月間のクローリング時には非公開ユーザでも退会ユーザでもなかったが，クローリング終了時に非公開設定を行ったユーザや Twitter を退会したユーザのことである．彼らは，4. 章で述べる API を用いた調査方法ではデータを取得できないため調査対象外とした．この結果を図 1 に示す．ここで，「その他」のユーザは 1000 人 *7 中 33 人となった．従って，実質分類不可能なユーザは全体の 3.3 %となり，網羅性の高さが示された．

4. 調査

4.1 概要

この調査の目的は，アイコン画像別にユーザ特性を調べることである．まず，アイコン画像別に 100 人のユーザを集めた．それらに対して取得したデータは，ユーザのフォロー数，フォロワー数，ツイート数，リツイート数，メンション数，URL 付きツイート数，ハッシュタグ付きツイート数，時間帯別ツイート数，時間帯別ツイート率，URL ドメイン数，被リツイート数である．これらは，Twitter REST API*8 を用いて 2013 年 10 月 14 日から 11 月 13 日の利用履歴を取得した．この調査結果と考察を以下に述べる．

*5 情報をフォロワー以外に公開していないユーザ
*6 Twitter サービスの利用を辞退したユーザ
*7 1067 人から非公開ユーザと退会ユーザを除いた数
*8 https://dev.twitter.com/docs/api/1.1

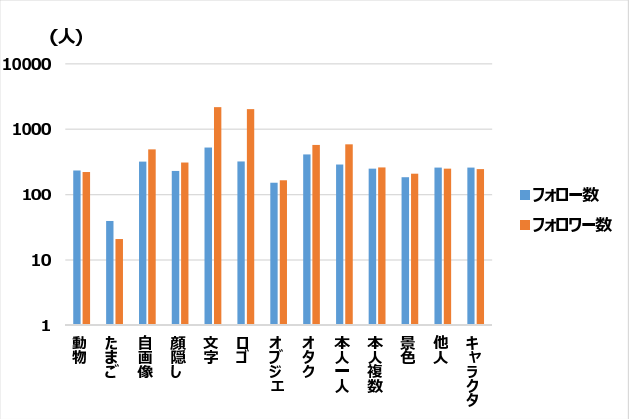

図 2: フォロー数・フォロワー数の中央値

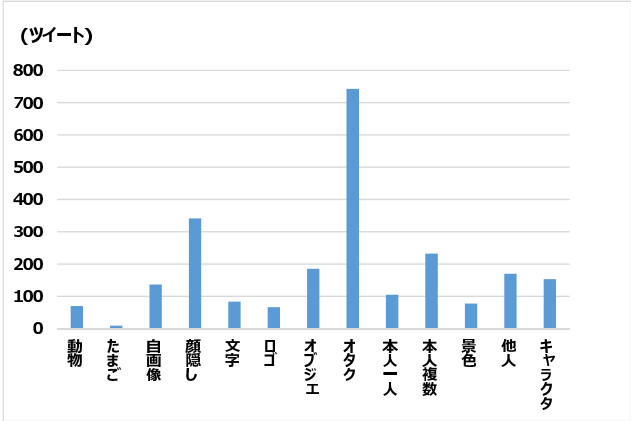

図 3: ツイート数の中央値

4.2 結果・考察

ここではスペースの都合上，フォロー数とフォロワー数，そしてツイート数に関する結果を挙げる．また，以下から，それぞれのアイコン画像に属するユーザを，動物アイコンに属するユーザは動物ユーザ，景色アイコンに属するユーザは景色ユーザなどと呼ぶことにする．

4.2.1 フォロー数・フォロワー数

図 2 は，それぞれのアイコン画像におけるユーザのフォロー数とフォロワー数の中央値を棒グラフにまとめたものである．この図において，縦軸が対数表示になっていることに注意されたい．例えば，他人ユーザにおいてフォロー数の中央値は 258.5，フォロワー数の中央値は 249.5，などと読み取れる．

図 2 から，フォロー数とフォロワー数において共に文字ユーザが最大値であることが分かる．文字ユーザには，非公式な bot[*9] や学生サークルのアカウント，個人営業を行うアカウントなどが多い．つまり，彼らは自身のアカウントに他のユーザの注目を集めようとするユーザである．彼らは，宣伝や広告を行うために，Twitter 上に存在する “フォロー返し ” の慣習 [13] を用いてフォロワーを獲得し，結果的にフォロー数もフォロワー数も最大値をとったと考えられる．

また，ロゴユーザにおいては，フォロー数に対してフォロワー数の値が大きい．ロゴユーザには，会社や企業などの組織を代表する公式アカウントが数多く存在する．例えば，「AmazonJP 本のお得情報」というアカウントは，e-commerce サイトの代表例である Amazon[*10] で販売される本の新刊情報を発信するアカウントで，フォロー数は 7 人だがフォロワー数は 33068 人である．彼らは，実社会ですでに有名である場合もあるため，ロゴユーザが発信する情報は Twitter ユーザからの注目度は高い．さらに，彼らのツイートの大半はフォロワーにとって有益な情報であるため，一度フォローしたユーザはフォローを外さない傾向にあることも要因の一つである．

たまごユーザは，フォロー数，フォロワー数共に最小値である．たまごの画像は，Twitter サービスを利用し始めて最初に標準で設定されている画像であるため，たまごユーザは典型的な初心者ユーザであることが分かる．

*9 Twitter の機能を使って作られた機械による自動発言システム (広義にはそれを真似た手動でツイートするユーザアカウントも含む)

*10 http://www.amazon.co.jp/

4.2.2 ツイート数

図 3 は，それぞれのアイコン画像におけるユーザのツイート数の中央値を棒グラフでまとめたものである．動物ユーザの中央値は 69.5，景色ユーザの中央値は 77.5 などと読み取れる．

この図から，オタクユーザのツイート数が目立って高いことが分かる．オタクユーザとは，ほとんどが美少女系アニメの画像，公共性の低いアニメやゲームの画像を使うユーザである．彼らは自身の趣味や嗜好に関する投稿が多く，その数もかなり多い．また，他のユーザとのやり取りを行うより，一方的な情報発信を行うユーザが多い．そのためにツイート数が多くなる傾向にあると考えられる．ただし，リツイート回数や Web アプリケーションの URL が引用されることも多いため，投稿テキストは趣味・嗜好に偏りがちである．

また，文字ユーザやロゴユーザはフォロワー数が大きい値であったのに対し，ツイート数は小さい値をとっている．彼らは主に宣伝や広告を行うユーザであると前述したが，宣伝・広告はその回数やコンテンツの影響を受けやすいと考えられる．同じ商品を短期間に何度も宣伝するよりは，様々な種類のコンテンツを宣伝されるほうがユーザには好まれやすいと考えられる上に，その商品の在庫がない場合にはその宣伝をすることは出来ない．つまり，自身のアカウントをより多くのユーザに注目してもらうためには，ただ数多く投稿するのではなく，様々なコンテンツを適度な回数で投稿する必要がある．これが，彼らのツイート数の少なさに繋がっている．また，基本的にはフォロワーとのやり取りを避ける傾向にあることも要因の一つであるといえる．

5. まとめ

今回の調査結果では，Twitter 上で 13 種類のアイコン画像を定義し，分類を行った．ここで，誰が分類しても一致するのか，13 分類項目でアイコン画像を網羅できるのか，という 2 点の検証を行った．その結果，Siegel の一致係数により複数人の被験者の分類は実質的に一致することが示され，統計的に十分なユーザ数を取得した後に 13 項目に分類することで網羅性の高さも示された．そして，以上の 2 点の検証により，それぞれのアイコン画像に属するユーザ 100 人に対し，各データを取得した．フォロー数，フォロワー数，ツイート数に関してその結果と考察を述べた．この調査結果から，アイコン画像による分類と取得したデータとの間に強い相関はないということが言える．確かに，文字ユーザのフォロー数やフォロワー数が

大きいことや，ツイート数が比較的小さいことなど，いくつか目立った差は発見された．しかし，すべてのアイコン画像の間に目立った差が見られたわけではない．これは，今回取得したデータのみでは，アイコン画像ユーザの行動特性を十分に推測できないということになる．一方で，今回の調査では，ユーザがアイコン画像を選択する際に影響する価値観や嗜好に関する詳細な調査は行っていないため，アイコン画像とユーザの間にある知見を心理学的側面から調査することを今後の課題としている．工学的には，ユーザクラスタリングやコミュニティ抽出など，ユーザとネットワークに着眼した応用を行う予定である．我々の仮説，アイコン画像はユーザの内面的特徴を示す可能性がある，ことが立証されれば，今回の調査結果を含め今後行う心理学的知見が工学的にも有用であると考えている．

参考文献

[1] 奥村学，“ソーシャルメディアを対象としたテキストマイニング” 電子情報通信学会 基礎・境界ソサイエティ Fundamental Review Vol.6 No.4 pp.285–293, 2013.

[2] 榊剛史，“ソーシャルセンサとしての Twitter -ソーシャルセンサは物理センサを凌駕するのか-” 人工知能学会誌 Vol.27 No.1 pp.67-74, 2012.

[3] Johan Bollen, Alberto Pepe, and Huina Mao, “Modeling public mood and emotion:Twitter sentiment and socio-economic phenomena”, In WWW2010, pp.450-453, 2010.

[4] Sitaram Asur and Bernardo A. Huberman, “Predicting the future with social media”, In WI2010, pp.492-499, 2010.

[5] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors”, In WWW2010, pp.851-860, 2010.

[6] Marco Pennacchiotti and Siva Gurumurthy, “Investigating topic models for social media user recommendation”, In WWW2011, pp.101-102, 2011.

[7] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li, “Comparing twitter and traditional media using topic models”, In ECIR2011, pp.338-349, 2011.

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, “Information credibility on twitter”, In WWW2011, pp.675-684, 2011.

[9] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei, “Rumor has it: Identifying misinformation in microblogs”, In EMNLP, pp.1589-1599, 2011.

[10] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, David Yarowsky, “Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter” In NAACL-HLT, pp.1010-1019, 2013.

[11] Marco Pennacchiotti and Ana-Maria Popescu, “A Machine Learning Approach to Twitter User Classification” In ICWSM, pp.281-288, 2011.

[12] 池田謙一，“ネットワーキング・コミュニティ” 東京大学出版，1997.

[13] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He “TwitterRank Finding Topic-sensitive Influential Twitterers” In WSDM , pp.261-270, 2010.