7Y-05

ソーシャルメディア上の行動データから流出する個人情報の定量的分析

畑田裕二 ⋆　矢谷浩司 †

東京大学工学部電子情報工学科 ⋆
東京大学大学院工学系研究科電気系工学専攻 †

1　はじめに

スマートフォンとソーシャルメディアの普及とともに，個人情報の流出経路が多様化している．Twitter などのソーシャルネットワーキングサービス（SNS）にはテキスト投稿だけでなく，他者と交流したり他者の投稿に対してリアクションを示したりするための機能が存在する．こうした SNS 上での行動データは，ユーザが意図的に発信しているつもりの無い個人情報を含んでいる可能性がある．本稿では Twitter 上でユーザが投稿したテキスト・リアクションした他者のツイート・フォローしているアカウントのプロフィール文に含まれる単語を集計し，それぞれのデータから流出する個人情報の重要度を算出した．その結果，本人が個人情報に関する投稿をしていない場合でも，フォローしているアカウントのプロフィール文から多くの個人情報を得られることが分かった．

2　関連研究

ソーシャルメディア上のデータを分析することで，個人の性格や行動の傾向を推定した先行研究は多く存在する．Pfeil et al [1] はコミュニティサイトの一つである MySpace*1ユーザの，性別・年齢といったデータや，プロフィール欄に記入されたテキストデータを収集して分析を行った．その結果，10 代の若年層ユーザと 60 歳以上のユーザでは，SNS で交流している相手の年齢分布などが異なることが明らかになった．他にも Twitter の投稿テキストから，アカウント所持者の性格を推定した研究も存在している [2]．このように，ソーシャルメディア上のデータを用いることで，本人は発信しているつもりの無い情報を得られる場合がある．

3　Twitter を用いた個人特定度の分析

本稿では，ソーシャルメディア上の行動データから個人情報がどの程度得られるか定量的に調査する．Twitter 上から，ユーザが意識的に発信している情報とそうでない情報を含む行動データを収集し，それらに含まれる個人情報を「個人特定度」という指標を用いて定量評価し比較を行う．

3.1　分析対象とデータ

分析に用いる行動データを収集するため，以下の条件を満たす Twitter アカウントを 10 人分選定した．

1. 我々と直接の面識があり，個人情報を把握している人物のアカウントであること
2. プライバシー設定が公開設定であり，誰でもそのアカウントの行動データを閲覧できること
3. そのアカウントからは既に 1000 ツイート以上の投稿がなされていること

これらのアカウントに対して，以下のテキストデータを Twitter REST API を用いて収集し，データセットを作成した．収集日時は 2017 年 12 月 12 日である．

- *OwnTweet*: 本人が行ったツイートと返信（1 名あたり最大 3000 件，計 21952 件)
- *Likes*: 本人が「いいね」を付与した他者のツイート (1 名あたり最大 3000 件，計 15869 件)
- *FollowProfiles*: 本人がフォローしているアカウントのプロフィール文 (計 4869 人分)

3.2　個人情報の定量評価指標

ある語が持つ個人情報としての重要度を「個人特定度」とする．この個人特定度は安井ら [3] を参考に，その個人情報が明らかになることで，母集団からどの程度個人を絞り込めるかに基づいて定めた．例えば，個人の所属大学が分かった場合，日本国内には大学・短大は約 1000 校存在することから，$1000 \simeq 2^{10}$ で 10 bit の個人特定度とした．大学の所属サークルや学科が明らかになった場合も，その大学に存在するサークルや学科の数に基づいて個人特定度を定めた．

他の情報と組み合わせて初めて母集団の人数が明らかになるものは，一旦 1 bit を与えておき，後で必要な情報が明らかになった時に再度個人特定度を求めて与えた．例えば野球部に所属していることが分かっても，日本に存在する課外活動団体の数が分からないので，「所属部の有無」という情報として 1 bit 与えておく．後に所属大学が明らかになった際に，大学内の課外活動団体数に基づいて個人特定度を置き換える．ただし，大学名が分からない状態で学部だけが明らかになった場合，大学に存在する学部を $8 = 2^3$ 個と仮定して個人特定度を与えた．その他の情報として，アルバイトをしているか否かやサークルの代，通学に使用している路線には 1 bit を与えた．

Quantitative Analysis of Personal Information Disclosure from Behavior Data on Social Media
Yuji HATADA⋆ and Koji YATANI†
⋆† Interactive Intelligent Systems Laboratory,
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{hatada, koji}@iis-lab.org
*1 https://myspace.com/

表1: 個人特定度の計算例（*FollowProfiles* に出現した上位 6 単語のみ抽出）．このように各アカウントの 3 つのデータセットについて出現頻度上位 100 単語の個人特定度を求め，それらを累積する．

出現単語	個人特定度	累計個人特定度	個人特定度の計算方法
野球部	1	1	サークルへの所属のみ判別.
東大	17	18	日本の大学数（10 bits）＋東大の課外活動団体数を 256 と仮定し，野球部の 1 bit を 8 bits に更新.
電情	5	23	学科数は 44.
工学部	0	23	学科により学部は特定済み.
五月祭	0	23	大学特有の学園祭だが大学特定済み.
4 年	2	25	学年は 4 種類.

3.3 分析手法

得られた 3 つのデータセットそれぞれを，新語を含む辞書 mecab-ipadic-NEologd[*2]を用いて品詞ごとに分かち書きし，名詞のみを抽出した．この時，BCCWJ 主要コーパス語彙表[*3]に記載されている上位 1000 単語を，一般的な語として取り除いた．その後，各人のデータセットごとに単語の出現頻度をカウントし，多い順に 100 個並べた．最後に個人情報とみなされる語に対して前節の手順で個人特定度を定め，それらを累計することで明らかになる個人情報の量を定量評価した．分析手順の例として，表 1に模擬データを用いて分析の様子を示した．

4 分析結果

10 人分の個人特定度の平均の推移を，図 1に示す．また 1 アカウントあたりの個人特定度の平均は，使用データセットごとに表2の通りであった．3 つのデータセットの中で，*FollowProfiles* を用いた場合が，最も個人特定度が高くなっている．また *Likes* による個人特定度は，3 つの中で最も標準偏差が大きかった．さらに分析を行った全員について，*FollowProfiles* を使用した場合に最も個人特定度が高い結果となった．これはすなわち，たとえユーザ自身が個人情報を含む呟きをしていなかったとしても，Twitter 上の行動（フォロー）から個人情報が流出する可能性があることを示唆している．

5 考察

FollowProfiles による個人特定度が最も高いのは，若年層の Twitter の使い方によるものと考えられる．Pfeil et al [1] は，10 代ユーザの MySpace の使い方は年の近い友人との交流が中心であることを報告している．若年層の SNS ユーザは友人などに自身を認知してもらうために，プロフィール欄に経歴や所属サークルなどを書くことが多いのだと考えられる．

Likes による分析の標準偏差が大きかったのは，Twitter の「いいね」機能の性質が影響していると考えられる．「いいね」機能は，本人の趣味や興味を反映したり友人とのコミュニケーションに用いられたりするため，必ずしも本人の個人情報が反映されるとは限らない．趣味など個人情報以外の話題に対する反応が活発であるほど，*Likes* から得られる個人情報は少なくなる．

*2 https://github.com/neologd/mecab-ipadic-neologd

*3 http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

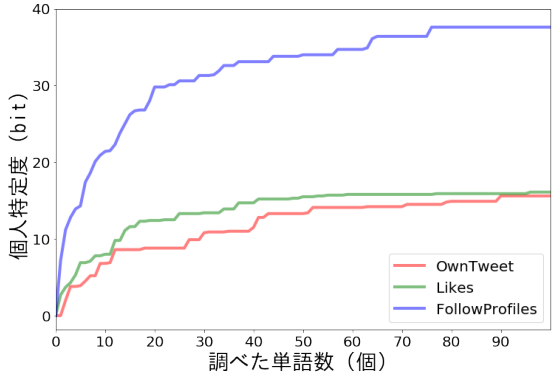

図1: 10 人分の個人特定度の平均の推移．データセット中に頻出した上位 100 単語を順に調べ，単語数ごとの個人特定度の累計を算出した．*FollowProfiles* を使用した場合が最も個人特定度が高くなる．

表2: Twitter の行動データにおける平均累積個人特定度．

データセット	平均累積個人特定度
OwnTweet	15.6 bits (SD=2.96)
Likes	16.1 bits (SD=3.26)
FollowProfiles	37.6 bits (SD=2.57)

6 おわりに

本稿では，Twitter において，本人が必ずしも意図していない経路で個人情報が流出する可能性について調査した．その結果，たとえ自身の投稿内容に気を付けていても，「誰をフォローしているか」という情報から個人特定に繋がる可能性があることが分かった．

今後の課題としては，分析するアカウントの数を増やすことや，個人特定度の基準作成・算出をより厳密に，そして機械的に行う手法を検討することが挙げられる．Web 検索を用いることで，ある単語が個人を特定するのにどれだけ有用か評価できる可能性がある．例えば「EEIC」という語を Web 検索すると，電気系の学科であるという情報に加えて，従属関係にある東京大学・工学部という情報も引き出すことができる．個人情報をアルゴリズム的に分析できるようになった場合，プライバシー等に関わる様々な問題が生じる．SNS の普及に伴って変化しつつある個人情報の扱い方について，改めて考える必要がある．

参考文献

[1] Pfeil, U., Arjan, R. and Zaphiris, P.: Age differences in online social networking - A study of user profiles and the social capital divide among teenagers and older users in MySpace, *Computers in Human Behavior*, Vol.25, No.3, pp.643–654 (2009).

[2] Golbeck, J., Robles, C., Edmondson, M. and Turner, K.: Predicting Personality from Twitter, Proc. *PASSAT and SocialCom '11*, pp.149–156, IEEE (2011).

[3] 安井良介，佐藤和紀，針谷友彰，金井敦，廣田啓一，谷本茂明：ブログにおける個人情報漏えいレベルの定量化，情報処理学会 EIP 研究会研究報告，2009-EIP-43，Vol.2009，No.11，pp.9–16（2009）．