スクリーンネームを用いた ユーザの投稿活動率の推定手法に関する一検討

武田 悠佑[†, a]　　山本 修平[††, b]　　佐藤 哲司[††, c]

† 筑波大学 情報学群 知識情報・図書館学類　†† 筑波大学 大学院図書館情報メディア研究科

a) *s1311524@klis.tsukuba.ac.jp* b) *yamahei@ce.slis.tsukuba.ac.jp* c) *satoh@ce.slis.tsukuba.ac.jp*

概要　ユーザは，自身の名前や趣味，属性に応じたキーワードを含んだスクリーンネームで Twitter に登録し，多様な投稿活動をしている．ここで，スクリーンネームの中には，ユーザの投稿活動を特徴付けるような文字列が存在すると考えられる．本研究では，ユーザのスクリーンネームを構成する文字列に基づき，ユーザの Twitter における投稿活動率を推定可能か検証する．大量のユーザのスクリーンネームに対して長期間の分析をした結果，ユーザの投稿活動率を推定できるような特徴的な文字列が確認できたので報告する．

キーワード　Twitter，スクリーンネーム，投稿活動率，自己相互情報量

1　はじめに

2006 年にサービスを開始した Twitter[1] は，ますますその利用の定着が社会的に進んでいる．ユーザが Twitter を利用する目的の一つに情報の獲得があり，その達成手段としてユーザは自身の興味に関する情報を発信する他のアカウントをフォローする．このため，ユーザの持つ属性や興味を推定し，そのユーザと類似した興味や属性を持つ新たなフォロー先ユーザを検出する研究が数多く行われている．Hannon ら [2] は，ユーザの過去のツイート内容と，自身をフォローするユーザ群であるフォロワーと，自身がフォローするユーザ群であるフォロウィーのリストを特徴に用いた，ユーザ推薦手法を提案している．Gong ら [1] は，ユーザの結婚の有無，支持政党，宗教などの属性を，1 週間あたりの投稿活動率でユーザ群を分割して推定している．

一方で，Myers ら [3] は，ほとんどツイートしない，あるいは，過度にツイートするアカウントは，フォローから外されやすいことを明らかにしている．このことから，各アカウントの投稿活動率をフォローする際の特徴の一つとしてユーザに提示することも有益であると考えられる．

ユーザは,「スクリーンネーム」と呼ばれる英数字とアンダーバーから構成される唯一のユーザ ID を，自身で考え Twitter に登録する．スクリーンネームは，他のユーザにリプライをする際に頻繁に使用されることから，ユーザの名前や興味，属性，投稿活動率に応じた特徴的なキーワードが含まれていると考えられる．

本研究では，スクリーンネームを用いてユーザの投稿活動率がどの程度推定できるか分析する．ユーザを投稿活動率に基づいてグループに分割し，各グループで出現しやすい，あるいは出現しにくいキーワードを相互情報量によって抽出する．

[1] https://twitter.com/

2　投稿活動率とスクリーンネームの相関分析

2.1　データセット

分析には，Twitter Search API [2] を使用して収集した，2012 年 5 月 1 日から 2013 年 4 月 30 日までの一年間に日本語で投稿されたツイートを用いる．このうち，少なくとも毎月にツイートを 1 回以上投稿した，3,031,453 ユーザを分析対象とする．

2.2　分析手法

本論文では，ユーザの週毎の投稿活動への参加時間を投稿活動率と定義する．すなわち，ユーザが単位時間の間に，ツイート，リプライ，リツイートを行っていれば，投稿活動に参加しているとする．ある週が 7 日間で，ユーザがある 1 時間のみツイートを投稿した場合，投稿活動率は $\frac{1}{7\times 24}$ となる．各ユーザの投稿活動率を収集期間 53 週毎に算出した後，その平均値 a について，以下の条件を満たすグループ n にユーザを分類する．

$$2^{-n} < a \leq 2^{-n+1}. \tag{1}$$

次に，全てのユーザのスクリーンネームから，部分文字列として文字 3-gram を抽出する．各 3-gram の投稿活動率への関連の強さを計るため，本論文では特徴選択において有効性の知られている自己相互情報量を用いる．各グループ毎に出現する文字 3-gram の頻度を数えることによって，グループと文字 3-gram のクロス集計表を作成し，自己相互情報量を算出する．各グループにおいて自己相互情報量の高い文字 3-gram は，そのグループに出現しやすい文字 3-gram，自己相互情報量の低い文

[2] https://dev.twitter.com/docs/api/1/get/search

表 1 グループ分けの概要

	投稿活動の目安となる参加時間	ユーザ数
G1	1 日あたり 12-24 時間	82,907
G2	1 日あたり 6-12 時間	299,335
G3	1 日あたり 3-6 時間	565,582
G4	1 日あたり 1.5-3 時間	668,361
G5	1 日あたり 0.75-1.5 時間	612,467
G6	1 週あたり 2.63-5.25 時間	457,124
G7	1 週あたり 1.31-2.63 時間	257,588
G8	1 月あたり 2.91-5.81 時間	80,698
G9	1 月あたり 1.45-2.91 時間	7,391
計		3,031,453

表 2 3-gram 収集結果の概要

	総出現回数	種類数
All	26,386,023	49,963
Over 1000	21,551,467 (81.68%)	4,412 (8.83%)

字 3-gram は，そのグループに出現しにくい文字 3-gram とみなせる．

2.3 分析結果

ユーザを平均投稿活動率によってグループ分けした結果を表 1 に示す．左列はグループ番号を表しており，投稿活動率の目安とともにユーザ数を示している．最も多くのユーザが属するグループ G4 のユーザが参加する時間は 1 日あたり 1.5 時間から 3 時間であり，その数は 668,361 であった．最も少ないユーザが属するグループ G9 のユーザが参加する時間は 1 日あたり 0.047 時間から 0.094 時間、すなわち平均して 1 日あたり 4.2 分であり，その数は 7,391 であった．なお G9 の範囲よりも平均投稿活動率が低いユーザが存在しないのは，少なくとも毎月にツイートを 1 回以上投稿したユーザのみを対象としたためである．

スクリーンネームから文字 3-gram を抽出した結果の概要を表 2 に示す．出現回数が 1000 以上の 3-gram の種類は全体の 10 %に満たないにも関わらず，その出現回数の合計は全体の 80 %を超えていた．

全ての 3-gram について各グループとの自己相互情報量を算出した．総出現回数が 1000 回以上の 3-gram の内で，各グループにおいて自己相互情報量が上位 5 件だったものと下位 5 件だったものをそれぞれ表 3 と表 4 に示す．

2.4 考察

表 3 と表 4 を見ると「bot」という 3-gram が，G1，G2 では上位 1 位として，G3，G4，G5 では下位 1 位，G6 では下位 3 位，G7 では下位 4 位として出現していることがわかる．これは「bot」という 3-gram を含むスクリーンネームを持つユーザは，比較的高頻度で定期的に投稿を繰り返す bot 機能を有しているためであると考えられる．また「aaa」という 3-gram に注目すると G3，G4 では上位 1 位である一方で，G1，G6，G7，G8，G9 という 5 つのグループで下位 1 位として出現していることがわかる．これは「bot」のような経験的に知られているスクリーンネームの部分文字列ではないものの，明らかにその出現に偏りがあり,「aaa」を含むスクリーンネームを持つユーザは，G3，G4 に属するようなユーザであると推測できると考えられる．

表 3 各グループの自己相互情報量上位 5 件の 3-gram

順位	G1	G2	G3	G4	G5	G6	G7	G8	G9
1	bot	bot	aaa	aaa	chi	shi	ian	ian	ian
2	_bo	_bo	ooo	ooo	shi	tak	ani	ani	lia
3	a_b	___	uuu	iii	tak	chi	ang	tri	ani
4	i_b	kur	xxx	cha	cha	osh	and	ind	tri
5	n_b	252	iii	chi	aka	ama	tri	ang	adi

表 4 各グループの自己相互情報量下位 5 件の 3-gram

順位	G1	G2	G3	G4	G5	G6	G7	G8	G9
1	aaa	chi	bot	bot	bot	aaa	aaa	aaa	aaa
2	chi	cha	_bo	_bo	_bo	_bo	ooo	ooo	chi
3	cha	shi	shi	ani	___	bot	_bo	uki	uki
4	han	han	asa	ang	a_b	ooo	bot	chi	ooo
5	ooo	iii	tak	ian	an_	xxx	uuu	iii	aki

3 おわりに

本研究ではスクリーンネームを用いてユーザの投稿活動率をどの程度推測できるか分析するため，ユーザのスクリーンネームの文字 3-gram を抽出し，投稿活動率に基づき分割したユーザのグループと各 3-gram との間で自己相互情報量を算出した．その結果，その文字列を含むスクリーンネームを持つユーザの投稿活動率の推定が可能であるような部分文字列の存在が示唆された．

今後の課題としては，具体的な投稿活動率の推定手法の考案や，スクリーンネームを用いたユーザの投稿活動率の推定がどの程度可能であるかという定量的な評価方法の検討が挙げられる．

参考文献

[1] Gong, W., Lim, E.-P. and Zhu, F.: Characterizing Silent Users in Social Media Communities, *Proceedings of the ICWSM2015*, pp. 140–149 (2015).

[2] Hannon, J., Bennett, M. and Smyth, B.: Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches, *Proceedings of the RecSys2010*, pp. 199–206 (2010).

[3] Myers, S. A. and Leskovec, J.: The Bursty Dynamics of the Twitter Information Network, *Proceedings of the WWW2014*, pp. 913–924 (2014).