7K-02

統計的手法に基づいたウェブサービスにおけるボット検出

岡本　大輝†　　宮﨑　太郎†　　後藤　淳†

NHK 放送技術研究所†

1. はじめに

ウェブ上のマイクロブログサービスである Twitter には，プログラム由来で動作をするボット(以下bot)と呼ばれるアカウントが存在する．一般的に bot の投稿は話題性が低く，Twitter上で話題を探索する上で邪魔になりやすい．botは投稿内容や動作原理などにおいて様々な種類が存在し，一部は検出を避けるために動作を変化・高度化させる[1]．そのため，機械学習による検出は困難かつ労力を要する．本稿では，botの投稿動作の統計的性質に着目し，ルールベースの検出手法を提案する．

2. 提案手法 1: LiPP

人間が手動で投稿をする際には，投稿の時刻を細かく意識することは殆どないと推測される．この場合，人間の投稿を複数観測すると，投稿の時刻(タイムスタンプ)の[秒]および[分]の値の出現はポアソン過程になっていると仮定できる．

一方で，botはプログラムに従って動作するため，一般的には動作に規則性がある．Twitterのbotにおいて，規則的な動作の最たるものが投稿行動である．例えば『毎時○○分に投稿する』，『△△の発生を確認したら投稿する』，等が挙げられる．いずれの場合においても botのプログラムは，一定時間の待機処理と投稿可否の条件分岐を繰り返しており，タイムスタンプの[秒]および[分]の値は一定の規則に従って出現すると仮定できる．

これら 2 つの仮定から，タイムスタンプの[秒]と[分]の出現パターンのポアソン過程らしさ(Likelihood as Poisson Process: $LiPP$)を計算し，$LiPP$ が高いほど人間らしさが高いと評価する．

タイムスタンプの[秒]の値が k となる回数を $X_{sec}(k)$，同[分]の場合を $X_{min}(k)$ とする．k のとり得る値は 60 パターンなので，$X(k)$ の期待値 $E(X)$ および分散 $V(X)$ は，

$$E(X)=\sum_{k=0}^{59}\frac{X(k)}{60}\ ,\ V(X)=\sum_{k=0}^{59}\frac{\big(X(k)-E(X)\big)^2}{60}$$

となる．理想的なポアソン分布においては期待値と分散の値が等しくなるので，それらの比を $R(X)=V(X)/E(X)$ として，$LiPP$ を以下のように定義する．

$$LiPP=R(X)exp\big(1-R(X)\big)$$

計算の素となる $X(k)$ は[秒]と[分]のそれぞれで定義できるため，$LiPP$ も[秒]と[分]の 2 種類を定義できる．

3. 提案手法 2: NiPP

意図的か偶発的かを問わず，botの投稿タイミングにランダムなゆらぎが生じると $LiPP$ は機能しなくなる．そこで，投稿行動の大局的な特性に着目する．具体的には『botは1日前や1週間前も現在と同じような投稿行動をしている』『人間は bot に比べ日ごと・週ごとに行動が変化しやすい』と仮定し，1つのアカウントの行動パターンの時間誤差を評価する．仮定に基づくと，時間誤差が大きいほど人間であると評価できる．

ある期間における投稿の回数を数える行為は，数式上は δ 関数の総和と積分で表現できる．時刻 T_k に k 番目の投稿をするアカウントがあるときに，時刻 u から $u+w$ までの間の投稿数 $A_w(u)$ は，

$$A_w(u)=\int_u^{u+w}\sum_{k=1}^{N}\delta(t-T_k)\,dt$$

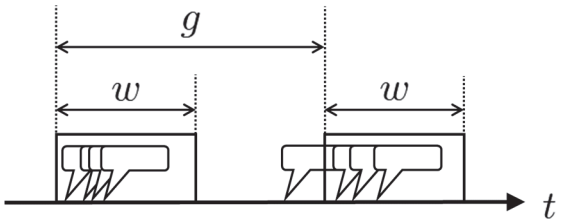
となる．ここで，N は対象とする全ての投稿の数である．

続いて，投稿回数の時間誤差を積分して正規化した値 $r_g(w)$ を定義する．さらに，$r_g(w)$ を引数として 0 以上 1 以下にスケーリングした値 $NiPP$ (Non-integer Period Power)を以下のように定義する．

$$r_g(w)=\frac{\int_{T_1}^{T_N-w-g}\{A_w(u)-A_w(u+g)\}^2du}{\int_{T_1}^{T_N-w-g}A_w(u)+A_w(u+g)\ \ du}$$

$$NiPP_g(w)=1-exp\Big(1-r_g(w)\Big)$$

ここで，g は時間誤差を評価する際に適用する時間差で，本稿では一貫して 24 時間とする．即ち，一つのアカウントで時間差 g (24 時間)，長さ w の 2 つの期間に投稿された投稿の数を比較して，時間誤差を算出する(図)．



図：投稿数の時間誤差算出のイメージ

時間軸(横軸)上の 2 つの窓に含まれる投稿数を比較する

4. 実験環境の構築

$LiPP$ と $NiPP$ の精度を評価する実験について述べる．

4.1. データセット

まず，Cresci らが公開しているデータ[2]を使用した．データに含まれる非botアカウント群に，botの中でも投稿数の多い 3 群(social spambot #1, #3, traditional spambot #3)を加えて，GIVEN DATASET とした．

また，Twitter社が提供するサービスを利用し，日本語で投稿される全ツイートの10%サンプルを入手して，その中からランダムに 200 アカウントを抽出した．それらのアカウントに対して“自動投稿に見えるか否か”で人手によるラベル付けをし，ORIGINAL DATASET とした．ここで，判別が難しいアカウントは除外した．

なお，GIVEN も ORIGINAL も 2018 年 7 月～9 月の 3 か月間に最低 1 回は投稿をしているアカウントのみを対象としている(表 1)．

BOT Detection on the Web Service Based on Statistical Method
†NHK Science and Research Laboratories

Copyright ©2019 Information Processing Society of Japan.
All Rights Reserved.

表 1:データセットの構成

	bot	非 bot
GIVEN	398	2146
ORIGINAL	63	102

4.2. アカウントの評価・分類手法

データセット内の全てのアカウントについて，Twitter の user_timeline API を利用して直近 200 ツイート(2018 年 9 月 25 日時点)を取得する．それらのツイートに含まれる情報を活用して，bot か否かを評価・分類する．

4.2.1. 提案手法

提案手法の $LiPP$ は，[秒]と[分]のそれぞれで出現回数を利用し，2 つの評価値 $LiPP_{sec}$ と $LiPP_{min}$ を個別に用いてアカウントを評価する．もう一つの提案手法 $NiPP$ では，g を 24 時間で固定した上で， w を 1 時間と 24 時間に設定した $NiPP_{hour}$ と $NiPP_{day}$ の 2 種類でアカウントを評価する．さらに，$LiPP$ と $NiPP$ の各 2 種類ずつ，計 4 つの評価値を全てかけあわせた値 LN を評価値とした評価も行う．

4.2.2. 比較手法

アカウントの投稿の時間間隔をもとにエントロピーを算出して乱雑さ(人間らしさ)評価する手法[3]と，アカウントの特徴量をサポートベクターマシン(SVM)で学習・分類する手法を用いる．SVM の学習に用いる特徴量としては，$LiPP$ の算出時に得られた期待値 $E(X)$ と分散 $V(X)$ それぞれ 2 種類ずつ，計 4 種類を用いる場合(SVM_{time})と，上記 4 種類に加えて user_timeline API で取得できるユーザ情報(表 2)を用いた場合(SVM_{u+t})の 2 種類とした．

表 2:SVM の学習に用いるユーザ情報

	数的性質
フォロー/フォロワー 比	正の小数
投稿数	自然数
アクティビティの有無 [プロフィル画像の使用 / リスト作成 / ツイートのお気に入り / 位置情報登録 / URL 登録]	0 or 1 (5 種類)

4.3. 精度検証の指標

ランキング精度の指標である AUC (Area Under the Curve)と，2 値分類精度の指標である MCC (Matthews Correlation Coefficient)を用いる．4.2.節に示したいずれの手法においても，データセットを 4:1 に分け，5-fold cross validation を 10 回行ってアカウントを評価または分類し，平均 AUC と平均 MCC で精度を検証する．

SVM による分類結果を AUC で評価する際には，テストセットに含まれる各アカウントの分離平面からの距離を評価値として，ランキングを作成する．

5. 実験結果と考察

実験結果を表 3 に示す．$LiPP$ と $NiPP$ をかけあわせた LN が概ね良好な精度を示している．

5.1. LiPP と NiPP の相乗効果

殆どの場合で， $LiPP$ 単体または $NiPP$ 単体より，組み合わせた LN の精度が優れている．両手法を信号処理的に解釈すると，$LiPP$ は高周波領域(短い時間スパン)に，$NiPP$ は低周波領域(長い時間スパン)に着目している．

高周波と低周波という逆の性質の特徴の組み合わせによって，効果的に精度が向上したと考えられる．

表 3:実験結果

	GIVEN		ORIGINAL	
	AUC	MCC	AUC	MCC
エントロピー	0.812	0.734	0.945	0.774
SVM_{time}	0.942	0.832	**0.991**	0.873
SVM_{u+t}	0.749	-0.004	0.574	0.000
$LiPP_{sec}$	0.911	**0.846**	0.951	0.867
$LiPP_{min}$	0.662	0.409	0.939	0.812
$NiPP_{hour}$	0.936	0.846	0.955	0.906
$NiPP_{day}$	0.905	0.698	0.949	0.853
LN	**0.954**	0.834	0.990	**0.912**

5.2. NiPP のパラメータ設定

$NiPP$ には g と w の 2 つのパラメータがある．予備実験として，w を 1 hour ($\therefore NiPP_{hour}$) から 24 hours ($\therefore NiPP_{day}$) まで変化させて精度を検証した．その結果，1 hour で最も精度が良く， 24 hours になるにつれて徐々に精度が劣化することが判明した．

5.1.節と同様に信号処理的な解釈に沿うと， $NiPP$ は『 w の時間間隔で生じる投稿頻度の変化』に対して大きな値を取りやすい．すなわち，人間の投稿行動はたかだか 1 時間程度しか継続せず，bot は人間に比べ長時間安定的に投稿する，と解釈することができる．この解釈は直観的にも納得しやすい．

5.3. SVM の性能と時間特徴量の寄与

実験に用いた 2 種類の SVM を比較すると， SVM_{time} の精度が顕著に高いことが分かる．SVM_{time} は，[秒]や[分]の数値の出現パターンの平均や分散という，きわめて原始的な特徴量を用いているが，それらが bot の検出に大きく寄与していることが示唆される．一方，SVM_{u+t} は学習に用いた特徴量の種類が多いにも関わらず，精度が劣化してしまっている．付加されたユーザ特徴量は bot と人を分類する効果が薄く，結果的にノイズとして作用してしまっていると考えられる．

6. まとめ

Twitter の bot アカウントを検出するために，投稿行動の統計的特徴に着目した手法を提案した．直近 200 投稿の時刻のみを利用し，高い精度で評価・分類できることを示した．本稿で提案した手法をユーザクラスタリングに応用し，より効率的な話題収集を実現していく．

参考文献

[1]. E. Ferrara et al., The Rise of Social Bots, COMMUNICATIONS OF THE ACM, Jul, 2016,

[2]. S. Cresci et al., The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race, IW3C2, Apr, 2017,

[3]. R. Ghosh et al., Entropy-based Classification of 'Retweeting' Activity on Twitter, SNA-KDD'11, Aug, 2011,

Copyright ©2019 Information Processing Society of Japan.
All Rights Reserved.