DBSJ Journal, Vol.12, No.1
June 2013

ベンフォードの法則を応用したbotアカウント検出

Bot account detection using Benford's law

蔵内 雄貴♡ 西田 京介◇ 川中 翔♠
星出 高秀♠ 内山 匡♠

Yuki KURAUCHI Kyosuke NISHIDA
Sho KAWANAKA Takahide HOSHIDE
Tadasu UCHIYAMA

Twitter アカウントの中には **bot** と呼ばれる自動生成アカウントが存在し，一般的なユーザを対象としたいマーケティングにおいてノイズとなっている．そのため，**bot** アカウントに特徴的な素性を利用し検出する手法が研究されているが，その精度は十分でない．そこで，従来利用されていなかった，単語頻度や投稿時間の分布をもとにした素性を利用することで，**bot** アカウントを高精度に検出する手法を提案する．具体的には，ベンフォードの法則に着想を得て，ユーザのツイートログに関する単語頻度と投稿間隔の数値集合における最上位桁の数値の出現確率を素性として分類器を構築することで，単語頻度と投稿間隔の分布を得るために大量のツイートが必要となる問題を解決する．実験では，**Twitter** データを用い，**bot** アカウントと **human** アカウントの判別タスクによって評価を行った結果，従来法で利用される素性群と比較して高い精度が得られることと，従来法と組み合わせることで精度が **9.2%**ポイント向上できることを確かめ，提案法の優位性が示された．

Twitter uses an automatic account response tool called bot; its output disturbs the value of tweet data for marketing since it targets general users. In response to this issue, bot detection through feature recognition has been attempted but its accuracy is still insufficient. We present a method of detecting bot accounts with high accuracy by using characteristic features based on the distribution of term frequency and posting time, in addition to the existing features. In detail, we overcome tweet scarcity in order to obtain the distribution of term frequency and posting interval. We propose a classifier whose features include are appearance probability of the highest-order digit in a number group of term frequency and posting interval of a user. This is inspired by Benford's law. To confirm the effectiveness of the proposed method, we conduct a two-class classification experiment; Twitter data is separated into bot accounts and human accounts. The proposed method offers higher accuracy than the conventionally used features. 9.2% improvement is achieved in a combination with the previous method.

♡ 正会員　日本電信電話株式会社 NTT サービスエボリューション研究所
kurauchi.yuki@lab.ntt.co.jp
◇ 正会員　NTT レゾナント株式会社　k-nishi@nttr.co.jp
本研究は，第二著者が日本電信電話株式会社在籍時に行われた．
♠ 非会員　日本電信電話株式会社 NTT サービスエボリューション研究所
{kawanaka.sho,hoshide.takahide,uchiyama.tadasu}@lab.ntt.co.jp

1. はじめに

情報の投稿・閲覧，コンテンツの共有手段であるソーシャルネットワーキングサービス（SNS）が普及し，様々な意見や感想が投稿されている．その中でも Twitter[1] サービスは，毎日 3 億件を越える投稿（ツイート）があり，かつ投稿内容の収集や解析が可能なことから，マーケティングのための情報源として注目が集まっている [19, 20].

Twitter アカウントの中には，プログラムによって自動的にツイートや友人登録（フォロー）などを行う bot と呼ばれるアカウントが存在する．マーケティングにおける対象ユーザは，消費者となる一般的なユーザであるため，これらの bot アカウントは対象ユーザとして不適切であり，除去すべきアカウントであると言える．しかし，プロフィールに bot であることを明示している bot アカウントもある一方で，明示していないものも多い．このような背景から，Twitter アカウントにおける bot アカウント検出が必要とされている．

従来の bot アカウント検出手法では，フォローしている人 (friend) とフォローされている人 (follower) の割合が大きく偏っている場合に bot である可能性が高いなどの，bot アカウントに特徴的な素性を用いて分類器を構築する方法が数多く提案されている [1, 5, 7, 8, 15]. しかし，これらの研究の精度にはまだ改善の余地がある．また，Chu ら [5] は，SNS におけるアカウントは，前述の bot アカウント，一般の人間である human アカウント，その 2 つの中間にあたる cyborg アカウントに分類できるとしているが，本稿では bot アカウントとは，Chu らの分類における bot アカウントと cyborg アカウントを含むものとする．

我々は，単語頻度と投稿間隔について，human アカウントと bot アカウントで従う分布が異なることに着目し，従来は利用されていなかった単語頻度と投稿間隔にもとづく素性を用いて分類器を構築することで，より高精度に bot アカウント検出を行うことを目指す．しかし，この際，単語頻度や投稿間隔の分布を得るためには大量のツイートが必要となるにも関わらず，bot アカウントの中には少量のツイートしか行わないアカウントがあり，正確な素性が得られない問題がある．このようなアカウントは，あるアカウントが多くのアカウントに支持されているように見せかけるなどのために，大量に作成されている．そこで，単語頻度や投稿間隔の最上位桁の数値の出現確率を素性として分布を判別することで，bot アカウント検出を行う手法を提案する．この手法で用いる素性である最上位桁の数値の出現確率は，9 次元であるため，少量のツイートからでも得ることができ，上記の問題が解決できる．これは，ある数値集合が対数的な分布に従って生成されている場合，その数値集合の各数値における最上位桁の数値の出現確率には偏りがあるというベンフォードの法則 [14] に着想を得たものである．

実験では，収集した Twitter データを用い，bot アカウントと human アカウントの判別タスクによって評価を行った．その結果，従来法で利用される素性群と比較して高い精度が得られることと，従来法と組み合わせることで精度が 9.2%ポイント向上できることを確かめ，提案法の優位性が示された．

2. 関連研究

bot アカウント検出は，マーケティングにおいて重要となる，トレンド検出 [10]，意見抽出 [6]，ユーザ属性推定 [11, 19] などの前

[1] http://twitter.com

表 1: 自然界における数値集合の，最上位桁の数値の出現確率の理論的な分布
Table.1: Theoretical frequency distribution of highest-order digit of number group in nature.

	1	2	3	4	5	6	7	8	9
出現確率 (%)	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

処理として必須の技術である．また，レコメンドにおいても，ユーザが扱った情報をレコメンドする場合 [3] などにおいて，その情報が bot アカウントから発信されていないかを判定する場合に有用である．

従来の bot アカウント検出手法では，bot アカウントに特徴的な素性を用いる方法が数多く提案されている [1, 5, 7, 8, 15]．用いられる素性には，総ツイート数，friend と follower の割合，投稿曜日の偏り，投稿ソース，URL を含むツイートの割合などがあり，推定器にはナイーブベイズ，ブースティングなどが用いられる．

類似のタスクである spam メール検出においては，spam で用いられやすい単語を利用する手法が研究されており [13]，spam ツイート検出にも適用されている [9]．これを適用することでソーシャルネットワークサービスにおける bot アカウント検出も可能であると考えられるが，そのように利用された例は見られない．これは，ツイート数が少ない bot も多く，このような場合に十分な単語量が得られず精度が低下するほか，トピックによって素性とすべき単語が変化する問題があるためだと考えられる．

別の類似のタスクであるユーザの性質を推定する研究としては，前述のユーザ属性推定 [11, 19] のほか，ユーザの影響力推定 [16] の研究があげられる．文献 [19] では，spam メール検出と同様に単語を利用するため適用できない．文献 [11, 16] では，ソーシャルネットワークをグラフとして用い，ユーザ属性や影響力を推定している．これらは，bot の周囲には bot が多いという性質や，bot 同士の特徴的なネットワークの形などを利用することで，同様に bot アカウントの検出が可能であると考えられる．これらの手法は，我々の手法を含む，特徴的な素性を用いた bot アカウント検出手法と組み合わせることで精度向上できると期待される．

我々の提案する手法は，bot アカウントに特徴的な素性を用いる手法に属し，従来利用されていなかった単語頻度と投稿間隔にもとづく素性を用いることで，その精度を向上することを目的とする．

3. ベンフォードの法則を応用した素性

本章では，まずベンフォードの法則について説明し，その後，ベンフォードの法則をもとにした素性と，Twitter におけるその素性の性質について述べる．

3.1 ベンフォードの法則の概要

ベンフォードの法則とは，電気料金の請求書，住所の番地，株価，人口，川の長さ，物理・数学定数などの，自然界に表れる（対数的な分布に従って生成される）数値集合の，各数値における最上位桁の数値の出現確率には偏りがあるという法則である [14]．表 1 に，その理論的な最上位桁の数値の出現確率の分布を示す．この分布から大きく離れている場合，なんらかの人手による操作などが行われた可能性がある．

この法則を利用すると，会計額の数値集合などが与えられた際に，その数値集合の各数値における最上位桁の数値の出現確率を求め，理論的な確率分布 (表 1) と比較することで，会計額が人手によって操作された可能性があることを検知できる．そのため，会計や支出に関する詐欺の指標として利用できることが示されている [12]．

3.2 素性としての最上位桁の数値の出現確率

我々はこのベンフォードの法則に着想を得て，最上位桁の数値の出現確率を素性とすることを考えた．そのメリットとして，与えられた数値集合の数が少なかったとしても分布に従うかを判別できることがあげられる．分布に従う度合を求める方法としては，他にも KL ダイバージェンスなどを用いる方法などが考えられるが，正確な分布を得るためには大量のデータが必要となる．例えば単語頻度分布などの場合，単語の数が極めて多いため，単語頻度分布を得るためには大量の文章が必要となる．しかし，最上位の数値の出現確率を用いる場合，素性とする最上位桁の数値は 1 から 9 のみであるため，最上位桁の数値の出現確率分布は少量のデータからでも十分な精度で得られる．すなわち，数値集合の分布の代わりに，最上位桁の数値の出現確率を用いれば，数値集合の完全な分布が得られないとしても，数値集合の分布と同等の情報量を得られることを，ベンフォードの法則は示唆している．

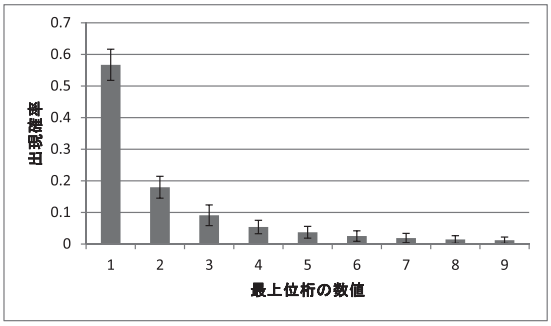

図 1: べき乗則（パレート分布）に従う乱数 (100 個) の場合
Figure.1: In case of 100 random numbers according to power-law (Pareto distribution).

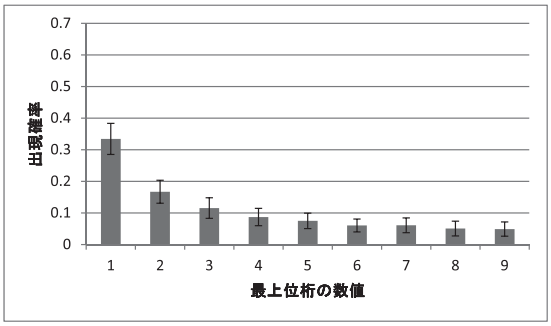

図 2: 指数分布に従う乱数 (100 個) の場合
Figure.2: In case of 100 random numbers according to exponential distribution.

そこで，数値集合の数が少ない場合に，最上位桁の数値の出現確率分布にどの程度の分散が生じるのかを，実験的に確かめた．具体的には，与えられた数値集合の数が 100 個であったと仮定して各分布に従う乱数を 100 個生成し，その最上位桁の数値の出現確率分布を求める，というステップを 100 回繰り返し，各数値の最上位桁の数値の確率と標準偏差を求めた．分布には，単語頻度が一般的に従う，べき乗則（パレート分布）と，投稿間隔が一般的に従う，指数分布を用いた．このとき，べき乗則（パレート分布）に従う乱数は式 (1)，指数分布に従う乱数は式 (2) によって生成した [18]．

$$(1-U)^{-1/\alpha}, \tag{1}$$

$$-\theta \ln(1-U), \tag{2}$$

ここで，U は $[0,1)$ の一様乱数，α は形状母数，θ は尺度母数を表し，$\alpha = 1$，$\theta = 1$ と設定した．

べき乗則の場合の結果を図 1 と表 2，指数分布の場合の結果を図 2 と表 3 に示す．図の横軸は最上位桁の数値，縦軸は各数値の出現確率，エラーバーは標準偏差を表し，表の列は最上位桁の数値を表す．その結果，標準偏差の最大値は，べき乗則における最

表 2: べき乗則（パレート分布）に従う乱数 (100 個) の場合
Table.2: In case of 100 random numbers according to power-law (Pareto distribution).

	1	2	3	4	5	6	7	8	9
mean	56.7	18.0	9.1	5.4	3.7	2.5	1.9	1.5	1.2
stdev	4.9	3.5	3.3	2.1	1.9	1.7	1.5	1.2	1.0
max	68.0	26.0	20.0	11.0	10.0	7.0	6.0	6.0	5.0
min	46.0	10.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0

表 3: 指数分布に従う乱数 (100 個) の場合
Table.3: In case of 100 random numbers according to exponential distribution.

	1	2	3	4	5	6	7	8	9
mean	33.4	16.7	11.6	8.7	7.5	6.0	6.1	5.1	4.9
stdev	4.9	3.6	3.2	2.7	2.5	2.0	2.3	2.3	2.3
max	55.0	27.0	24.0	19.0	16.0	12.0	12.0	12.0	12.0
min	22.0	6.0	4.0	4.0	2.0	2.0	1.0	0.0	1.0

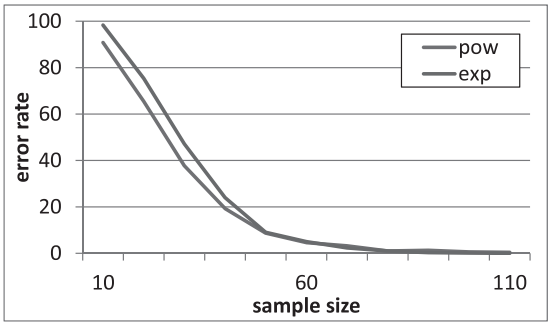

図 3: べき乗則，指数分布におけるエラー率
Figure.3: Error rate when according to power-law and exponential distribution.

上位桁の数値が 1 の場合で，4.9 であった．これは極めて小さく，例えば単語頻度を利用する場合，100 語の頻度が得られればその生成分布が何であるか判別が可能ということを示唆する．

さらに，与えられた数値集合の数による判別精度の変化を，検定を用いて検証した．具体的には，各分布に従う乱数から最上位桁の数値の出現確率を観測値として得た後，理論的な分布を期待値として，カイ二乗適合度検定 ($p < 0.05$) を行った．乱数の数は 10 から 500 まで 10 きざみとして検証した．そして，べき乗分布と指数分布について，それぞれ 1000 回実施し，エラー率を求めた．

結果を，図 3 に示す．横軸は与えた数値集合の数，縦軸はエラー率を表す．すなわち，図 3 の pow はべき乗則に従う数値を入力した際にべき乗則に従わないと誤った割合を，exp は指数分布の場合を示す．これから，いずれの場合も，数値集合の数が 100 程度のときにエラー率がほぼ 0 となることがわかる．こちらにおいても，先ほどと同じく，例えば単語頻度を利用する場合，100 語の頻度が得られればその生成分布が何であるか判別が可能という示唆を得られた．

3.3 素性としての最上位桁の数値の出現確率

ここでは，Twitter における数値集合として，単語頻度と投稿間隔を取り上げ，その最上位桁の数値の出現確率の検証結果を示す．

3.3.1 単語頻度

一般的に，単語頻度は zipf の法則に従い，k 番目に頻度の多い単語の頻度は最も頻度の多い単語の $1/k$ であることが知られている [17]．これは，Twitter において利用される単語の頻度においても同様であると考えられる．すなわち，Twitter における単語頻度の最上位の数値の出現確率は，べき乗則（パレート分布）に従う乱数の場合（図 1）に近いと予測される．

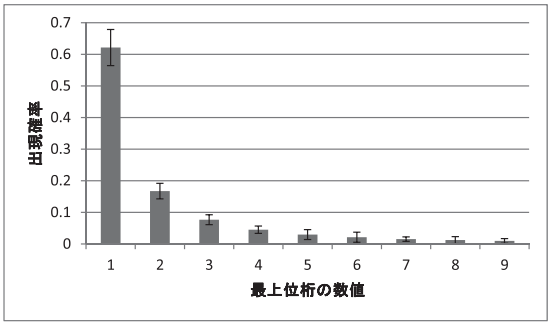

図 4: human アカウントの単語頻度の場合
Figure.4: In case of word frequency of human accounts.

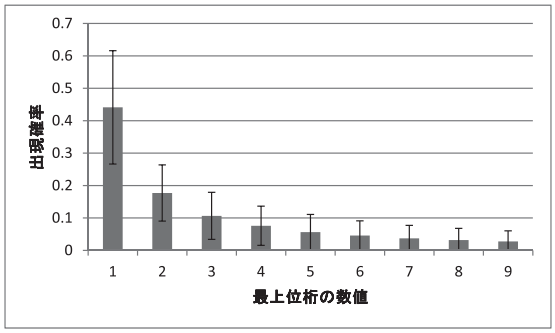

図 5: bot アカウントの単語頻度の場合
Figure.5: In case of word frequency of bot accounts.

しかし，bot アカウントの場合，投稿内容が特定のトピックに偏ったり，テンプレート文書を含むなどのため，一部の単語を集中的に利用する傾向があり，zipf の法則と図 1 には従わないと予想される．一方で，human アカウントの場合は，様々なトピックについて，テンプレートを持たず投稿するなどのため，多様な単語を用いる傾向があり，zipf の法則と図 1 に従うと予想される．

これを，実データを用いて実験的に確かめた．human アカウントの単語頻度の場合の結果を図 4 と表 4，bot アカウントの単語頻度の場合の結果を図 5 と表 5 に示す．これから，human アカウントはべき乗則（図 1）に従っており，標準偏差も最大で 5.7 と小さいことがわかる．bot アカウントはべき乗則（図 1）よりも指数分布（図 2）に近い特徴を持ち，標準偏差も最大で 17.5 と大きいことがわかる．

human アカウントにおいて，最上位桁の数値が 1 である確率が，べき乗則の場合よりも 5%程度高くなっている．これは，Twitter 特有の単語，すなわち，URL やハッシュタグなどが影響し，頻度が 1 となるような単語が多いためと考察できる．また，bot アカウントの標準偏差が大きいのは，bot アカウントにも様々な種類があり，全てのツイートがテンプレートに沿うような bot アカウントの場合，単語頻度がほぼ一様分布になるのに対し，広報などのアカウントではトピックに偏りがあるものの，human アカウントに近い分布をとることが影響していると考察できる．

3.3.2 投稿間隔

一般的に，単位時間ごとに生起する事象の数がポアソン分布に従う場合，そのような事象の生起間隔は指数分布に従う．これは，Twitter におけるツイートの投稿間隔においても同様であると考えられる．すなわち，Twitter における投稿間隔の最上位の数値の出現確率は，指数分布に従う乱数の場合（図 2）に近いと予測される．

しかし，bot アカウントの場合，投稿がプログラムなどによっ

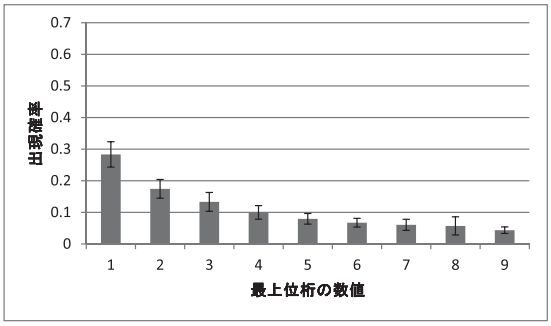

図 6: human アカウントの投稿間隔の場合
Figure.6: In case of time span of human accounts.

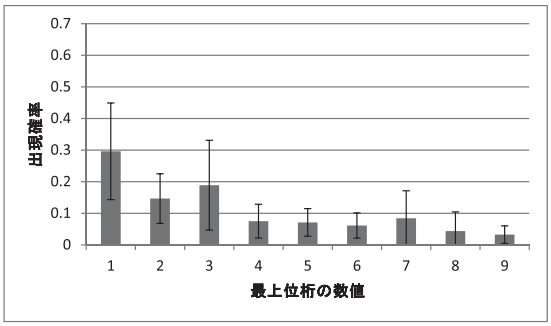

図 7: bot アカウントの投稿間隔の場合
Figure.7: In case of time span of bot accounts.

て制御され，投稿時間のルールや周期性を持つなどのため，投稿間隔がばらつかず固定される傾向があり，指数分布と図 2 には従わないと予想される．一方で，human アカウントの場合は，睡眠を含む様々な行動の合間に投稿を行うため，投稿間隔にばらつきが生じる傾向があり，指数分布と図 2 に従うと予想される．

これを，実データを用いて実験的に確かめた．投稿間隔は，秒単位として数値を得た．human アカウントの投稿間隔の場合の結果を図 6 と表 6，bot アカウントの単語頻度の場合の結果を図 7 と表 7 に示す．これから，human アカウントは指数分布（図 2）に従っており，標準偏差も最大で 4.0 と小さいことがわかる．bot アカウントは指数分布（図 2）に従っておらず，いくつかのピークを持ついびつな分布であり，標準偏差も最大で 15.3 と大きいことがわかる．また，bot アカウントにおいて特に 3 や 7 が大きな出現確率となっているが，これは，1 時間 (3600 秒) または 2 時間 (7200 秒) に 1 度投稿するような bot アカウントが多く存在することが影響していると考察できる．

以上から，単語頻度と投稿間隔ともに，human アカウントと bot アカウントにおいて，最上位桁の数値の出現確率の傾向が異なっていることがわかる．そのため，単語頻度と投稿間隔における最上位桁の数値の出現確率を素性として用いることで，human アカウントと bot アカウントの判別を行うことができると考えられる．

4. 提案法

ここでは，提案法の問題定義と概要について述べる．

4.1 問題定義

あるアカウントのツイート集合を入力とし，bot アカウントと human アカウントのいずれであるかを判別し出力する．この際，ツイート集合にはその投稿時間が付与されているものとする．

4.2 概要

あるアカウントのツイート集合を入力とし，各単語の頻度を数え上げた後，頻度の最上位桁における各数値の出現確率を求め，これを 9 次元の特徴ベクトルとした．この際，単語には，URL やハッシュタグを含む全ての品詞を利用し，ツイート中に含まれる短縮 URL は，再帰的に拡張した．投稿間隔についても，同様に 9 次元の特徴ベクトルとした．この際，投稿間隔の単位は秒とした．

この 2 種類の 9 次元特徴ベクトルは，重みづけなどを用いて統合するなど，様々な方法が考えられる．今回は，最もシンプルな方法としてそのまま結合し，18 次元の特徴ベクトルとして用いるものとした．

表 4: human アカウントの単語頻度の場合
Table.4: In case of word frequency of human accounts.

	1	2	3	4	5	6	7	8	9
mean	62.2	16.7	7.7	4.5	3.0	2.1	1.6	1.2	1.0
stdev	5.7	2.5	1.6	1.2	1.6	1.6	0.7	1.1	0.7
max	95.5	69.7	66.7	28.9	73.8	68.3	18.8	58.9	27.0
min	7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

表 5: bot アカウントの単語頻度の場合
Table.5: In case of word frequency of bot accounts.

	1	2	3	4	5	6	7	8	9
mean	44.1	17.7	10.7	7.6	5.6	4.6	3.7	3.2	2.8
stdev	17.5	8.7	7.3	6.0	5.4	4.5	4.0	3.6	3.3
max	98.8	75.4	81.2	71.8	80.8	66.4	72.8	72.7	50.7
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

表 6: human アカウントの投稿間隔の場合
Table.6: In case of time span of human accounts.

	1	2	3	4	5	6	7	8	9
mean	28.3	17.4	13.3	10.0	8.0	6.7	6.1	5.7	4.4
stdev	4.0	3.0	3.0	2.1	1.7	1.4	1.8	2.9	1.0
max	90.1	91.9	71.3	61.7	52.4	24.3	51.3	92.1	19.5
min	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0

表 7: bot アカウントの投稿間隔の場合
Table.7: In case of time span of bot accounts.

	1	2	3	4	5	6	7	8	9
mean	29.6	14.7	18.9	7.5	7.1	6.1	8.4	4.4	3.3
stdev	15.3	7.8	14.2	5.4	4.4	4.0	8.7	6.1	2.8
max	99.1	89.3	90.6	79.7	42.3	73.2	76.5	92.3	53.4
min	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0

5. 実験

提案法の有効性を検証するため，bot アカウントと human アカウントの分類タスクで評価実験を行った．本稿では，従来法で用いられた素性群と比較して，提案する素性がどの程度有用な素性であるのかと，従来法および提案法の 2 つの素性を組み合わせることでどの程度精度が向上するのかの 2 点を明らかにする．

5.1 データ

アカウントの収集方法は，下記の通りである．まず，bot アカウントについては，“bot”, “広報”, “PR” などの文字列を用いてプロフィールを検索し，該当したアカウントのプロフィールとツイートについて人手によるチェックを行い，bot アカウントと思われるものを収集した．次に，human アカウントについては，ランダムサンプルを行った後，各アカウントのプロフィールとツイートについて人手によるチェックを行い，human アカウントと思われるものを収集した．その後，各アカウントの過去ツイートを収集した．ツイートの収集には Twitter statuses/user_timeline API を用いた．収集したアカウント数はそれぞれ 5000 アカウントずつで，ツイート総数は 15,942,560 ツイートとなった．

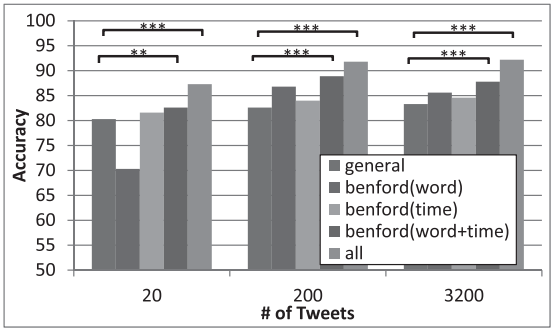

図 8: 実験結果 (Accuracy)
Figure.8: Result of experiment (Accuracy).

表 8: 定性的な実験結果の考察
Table.8: Qualitative discussion of experimental result.

アカウント種	単語頻度	投稿間隔	出力
一部に自動投稿を含む human	human	human	human
全てが自動投稿の human	bot	human	bot
リプライ bot（@recipetter）	bot	human	bot
検索キーワード反応 bot（@dororich）	bot	bot	bot
ツイート引用 bot（@1000Retweets）	human	bot	bot
ツイート生成 bot（@shuumai）	human	bot	bot

5.2 実験概要および結果

入力とするツイート数は，最大 20 件，200 件，3200 件のときの場合について実験を行った．推定器には，liblinear[2]の L2-regularized logistic regression (primal) を利用した．評価には，10-cross validation を用い，その際のパラメタは，予備実験においてテストデータとは別のデータセットを用意し，最適な値を求めてこれを利用した．評価指標には，accuracy を用いた．

比較手法には，[1, 5, 7, 8, 15] など多くの手法に共通して用いられている素性として，総ツイート数，friend と follower の割合，投稿曜日の偏り，投稿ソース，URL を含むツイートの割合を用い，提案法と同一の推定器によって判別する手法を用いた．

結を図果 8 に示す．図における横軸の数値は入力としたツイート数を，general は比較手法を，benford(word) は単語頻度の最上位桁の数値の出現確率を素性にした場合を，benford(time) は投稿間隔の最上位桁の数値の出現確率を素性にした場合を，benford(word+time) は上記 2 つの素性を組み合わせた場合を，all は全ての素性を組み合わせた場合を表す．

まず，入力とするツイート数の変化による精度の変化は，全ての素性において，ツイート数が多くなるほどに精度が上昇するが，200 ツイート以降は大きな変化がないことがわかる．また，最もツイート数の影響を受けるのは単語頻度 (benford(word)) であることがわかる．単語頻度と投稿間隔を組み合わせた場合 (benford(word+time))，ツイート数によらず，単体で用いた場合 (benford(word), benford(time)) よりも精度が上がっていることがわかる．これから，ツイート数が少ない場合でも，組合せる際に精度への悪影響を与えることはないと考えられる．

十分にツイート数がある場合には，単語頻度 (benford(word)) および投稿間隔 (benford(time)) は単一の素性のみを用いているにも関わらず，比較手法 (general) よりも高い精度が得られていることがわかる．特に，単語頻度 (benford(word)) のみを用いた場合には，86.8%の精度が得られており，比較手法 (general) の精度である 82.6%と比べても，4.2%ポイントの精度改善ができている．また，全ての素性を組み合わせる (all) ことで，91.8%と，比較手法 (general) よりも 9.2%ポイントの精度改善ができた．

6. 考察

human アカウント，bot アカウントにはそれぞれ様々な種類があるが，その内の典型的なアカウント種について bot を検出できているかを，定性的に考察する．典型的な種類として，一部に自動投稿を含む human，全てが自動投稿の human，リプライ bot，検索キーワード反応 bot，ツイート引用 bot，ツイート生成 bot を例にあげる．一部に自動投稿を含む human とは，foursquare などのユーザの代わりに投稿を自動で行うサービスを利用しているが，自らもツイートするようなユーザである．全てが自動投稿の human とは，自動投稿サービスを利用するが，自らはツイートをしない閲覧専門のユーザである．リプライ bot とは，ユーザのメンションツイートに対してリプライを行うような bot である．検索キーワード反応 bot とは，検索 API などを用いてユーザが特定のツイートをした場合にメンションツイートを行うような bot である．ツイート引用 bot とは，human アカウントが行ったツイートのうち，人気のあるツイートを紹介するような bot である．ツイート生成 bot とは，文章群から学習しツイートを自動生成するような bot である．

これらのアカウント種について，それぞれ単語頻度，投稿間隔が human，bot のいずれに近い分布をとるか，そして分類器の出力結果を表 8 にまとめる．単語分布について見ると，human アカウントでも，全てのツイートが自動投稿である場合には，bot に近いものとなる．一方，bot アカウントでも，ツイート引用やツイート生成を行う場合には，human に近いものとなる．投稿間隔について見ると，bot アカウントでも，ユーザの行動（チェックインなど）やメンションツイートに連動して投稿が行われる場合，human に近いものとなる．しかし，human アカウントのツイートを利用する bot でも，特定の単語や人気ツイートを検索した後に投稿する bot の場合は，機械的なタイミングで検索が行われた後に投稿が行われるため，bot に近いものとなる．このように，多くの bot アカウントにおいて，単語頻度もしくは投稿間隔のいずれかが bot らしさを持つため，正しく判別が可能である．しかし，human アカウントのうち bot に近い行動をとるユーザについては，判別が難しく，精度を下げる要因となったと考えられる．

提案法は，数値集合の最上位桁の数値の出現分布という，極めて一般的な素性を用いている．そのため，Twitter における bot アカウント検出だけでなく，様々なタスクに利用できると考えられる．例えば，ソーシャルネットワーク全般における bot アカウント検出や，ゲームなどにおける bot アカウント検出 [4]，spam メール検出，spam サイト検出などがあげられ，極めて一般的に異常検出として用いることができると考えられる．

ただし，ユーザ単位などのある程度のまとまりがないと最上位桁の数値の出現分布を構成できないために，ごく少量の文章などには適用が難しいという特徴を持つ．そのため，単一のツイートなどに対する処理，例えば，デマツイート検出などのタスク [2] などには適さないと考えられる．

従来の spam/bot アカウント検出の手法は，用いる素性が spam/bot 提供者側に知られてしまうと，spam/bot 提供者が対応し検出されないよう工夫されてしまう，いわゆるイタチごっことなってしまう問題があった．しかし，単語頻度や投稿間隔がべき乗則に従うようにツイートを生成することは難しいと考えられる．そのため，単語頻度や投稿間隔を素性にしていることが spam/bot 提供者側に知られても，対応されない，もしくは対応が難しいために，提案法にはイタチごっこを防ぐ効果も期待できる．

7. まとめ

本稿では，ベンフォードの法則を応用した bot アカウント検出手法を提案した．これは，従来利用されていなかった，使用単語や投稿時間の分布をもとにした素性を利用することで，bot アカウントを高精度に検出する手法である．本研究の貢献は，ベンフォード

[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

の法則に着想を得て，ユーザのツイートログに関する単語頻度と投稿間隔の数値集合における最上位桁の数値の出現確率を素性として分類器を構築することで，各分布を得るために大量のツイートが必要となる問題を解決し，bot アカウント検出の精度を改善した点にある．実験では，Twitter データを用い，bot アカウントと human アカウントの判別タスクによって評価を行った結果，従来法で利用される素性群と比較して高い精度が得られることと，従来法と組み合わせることで精度が 9.2%ポイント向上できることを確かめ，提案法の優位性が示された．

[文献]

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, volume 6, 2010.

[2] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[3] J. Chen, R. Nairn, and E. Chi. Speak little and well: recommending conversations in online social streams. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 217–226. ACM, 2011.

[4] K. Chen, H. Pao, and H. Chang. Game bot identification based on manifold learning. In *Proceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games*, pages 21–26. ACM, 2008.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 21–30. ACM, 2010.

[6] L. Dey and S. Haque. Opinion mining from noisy text data. *International journal on document analysis and recognition*, 12(3):205–226, 2009.

[7] D. Gayo-Avello and D. Brenes. Overcoming spammers in twitter-a tale of five algorithms. In *Spanish Conference on Information Retrieval (CERI)*, 2010.

[8] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM, 2010.

[9] D. Irani, S. Webb, C. Pu, and K. Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[10] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, pages 1155–1158. ACM, 2010.

[11] A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

[12] M. Nigrini. I've got your number. *Journal of Accountancy*, 187(5):79–83, 1999.

[13] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, pages 98–105. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998.

[14] H. Varian. Benford's law. *The American Statistician*, 26(3):65–6, 1972.

[15] A. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10. IEEE, 2010.

[16] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

[17] G. Zipf. The psycho-biology of language. 1935.

[18] 四辻哲章. 計算機シミュレーションのための確率分布乱数生成法. プレアデス出版, 2010.

[19] 蔵内雄貴, 内山俊郎, and 内山匡. マルコフ確率場を用いたソーシャルネットワークからのユーザ属性推定. *Web DB フォーラム 2012* 論文集, 2012.

[20] 池田和史, 服部元, and 松本一則. マーケット分析のための twitter 投稿者プロフィール推定手法 (コンシューマ・デバイス & システム vol. 2 no. 1). 情報処理学会論文誌 論文誌トランザクション, 2(1):82–93, 2012.

蔵内 雄貴　**Yuki KURAUCHI**
日本電信電話株式会社 NTT サービスエボリューション研究所研究員．2010 年慶應義塾大学大学院修士課程修了．同年日本電信電話株式会社入社．音楽情報処理，データマイニングの研究開発に従事．日本データベース学会会員．

西田 京介　**Kyosuke NISHIDA**
NTT レゾナント株式会社．2004 年北海道大学工学部情報工学科卒業．2006 年同大学大学院情報科学研究科修士課程修了．2008 年同博士課程修了．同年日本電信電話（株）入社．2013 年より現職．Web マイニングの研究開発に従事．博士（情報科学）．情報処理学会，電子情報通信学会，日本データベース学会各会員．

川中 翔　**Sho KAWANAKA**
日本電信電話株式会社 NTT サービスエボリューション研究所研究員．2009 年東京大学大学院修士課程修了．同年日本電信電話株式会社入社．データマイニング, 検索インターフェイスの研究開発に従事．

星出 高秀　**Takahide HOSHIDE**
日本電信電話株式会社 NTT サービスエボリューション研究所主任研究員．1993 年九州大学大学院総合理工学研究科修士課程修了．同年日本電信電話株式会社入社．遠隔教育システム，Web マイニングの研究開発に従事．情報処理学会正会員．

内山 匡　**Tadasu UCHIYAMA**
日本電信電話株式会社 NTT サービスエボリューション研究所主幹研究員．1987 年名古屋大学大学院理学研究科修士課程修了．同年日本電信電話株式会社入社．1998-2001 年 NTT コミュニケーションズ，2004-2006 年 NTT レゾナントにてポータルサービスの開発等に従事．2007 年より現職．行動モデリングの研究開発に従事．情報処理学会，電子情報通信学会，日本応用数理学会各会員．