2B4-NFC-02a-7

情報量に基づく投稿活動定量化手法を用いた東日本大震災前後のTwitter利用者の特徴付け

Characterizing Twitter Users around the Great East Japan Earthquake by Using Quantification Method of Posting Activities with Entropy

*1松本 慎平 Shimpei Matsumoto　*2川口 大貴 Hiroki Kawaguchi　*3鳥海 不二夫 Fujio Toriumi

*1広島工業大学情報学部
Faculty of Applied Information Science, Hiroshima Institute of Technology

*2広島工業大学大学院工学系研究科
Graduate School of Science and Technology, Hiroshima Institute of Technology

*3東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

At the time of the Great East Japan Earthquake, many Tweets of the disaster had posted and Twitter had been effectively-utilized as an infrastructure for sharing disaster information and confirming safety. However in Twitter, there have been various kinds of information and also the volume is extremely huge, so some kind of filtering mechanism to easily catch desired information is assumed to be needed for getting better performance out of Twitter at the time of disaster. To archive this, first of all characteristics of people actively posting disaster information should be grasped. This paper quantifies Twitter users on the basis of each user's posting activities, and analyzes the characteristics according to user's attributes. The real Twitter data distributed around the time of the earthquake is used, and especially in this paper, the difference between bot and human is mainly examined by using this data. This paper uses entropy to quantify user's activities.

1. はじめに

Twitterでは，実時間性の高い情報が各利用者から発信され，常時流通している．インターネット利用者は，これら情報を容易に閲覧できる．通常のインターネット掲示板と比較して，Twitterでは現在の出来事や場所に関する情報が多い．2011年3月11日の東日本大震災発生時，Twitterでは震災に関する多くの発言が行われた．多くは安否の確認や震災情報の拡散を目的に活用され，Twitterは情報インフラの1つとして，また有益な情報通信手段として機能した[風間 12, 鳥海 13]．被災地の状況を知る人間の投稿などマスメディアが報じない貴重な情報が数多く流通したことから，災害に関する情報を効率的に獲得する技術は有益であると考えられる．実際，Twitterの提供するサービスは災害時に有効であろうと考えられている[総務省消防庁 13]．しかしながら，Twitter上には様々な種類の投稿が膨大な量存在している．災害時にTwitterを有効に活用するためには，不必要な情報や目的以外の投稿をフィルタリングする仕組みが必要である．災害時に不必要な情報を的確に選別するためには，各利用者はどのような目的のもとでTwitterを利用しているのか，またどういった内容の投稿が流通しているのかを確認し，利用者を何かしらの指針に基づいて分類する必要があると考える．その準備段階として，本研究では，災害時におけるTwitter各利用者を投稿活動に応じて特徴付け，利用者それぞれを分類する手法の有効性を調査する．本研究では，震災前後実際にTwitterで流通した投稿を用いて分析を行った．とりわけ本研究では，利用者の分類を行いながら，bot利用者の特徴付けに重点を置いた．

連絡先: 松本慎平，広島工業大学情報学部知的情報システム学科，〒731-5193 広島市佐伯区三宅2-1-1，五日市キャンパス新4号館319号室，TEL/FAX: 082-921-6924，E-Mail: s.matsumoto.gk@cc.it-hiroshima.ac.jp

2. 解析手法の設計指針

従来の情報フィルタリング手法について考えてみると，一般的にメール分類やウェブページのコンテンツフィルタに利用されている．これらは情報の内容比較に基づく手法である．しかし，投稿内容だけで利用者とbotの分類を行うと，利用者のように振る舞うbotの分類が非常に困難である．従来のテキストによるフィルタリング手法は，静的な情報を対象とした質的解析である．一方Twitterは実時間性が高い動的なテキスト情報であるため，テキストに時間制が加わっていることが従来の手法だけでは十分ではない原因でないかと考えられる．そこで，情報量に着目し，Tweetが，テキストの質的な情報量を持つだけではなく，時間的尺度においても情報量を持っているのではないかと仮定した．投稿という行為そのものは従来のブログと同様にもかかわらずTwitterが注目を集めた背景には実時間性があり，また，人が実時間性に価値を見出した理由は，そこに情報量が存在していたからだと考えている．よって，情報量は，質だけではなく，投稿行動そのものにもあるのではないかとの仮定のもとで，時間に関する情報量に基づいてbotを評価すれば，利用者のように振る舞うbotの情報量は利用者に対して有意な差が見られるのではないかと考えた．

3. 解析手法

まず，投稿者は，5種類のエントロピーを利用することにより定量化される．本研究は，Ghoshらの手法[Ghosh 11]を拡張して利用した．エントロピーを時間間隔エントロピーと利用者エントロピーの2種類を定義する．これら2種類のエントロピーはそれぞれの扱う確率が違っている．

解析対象の全TweetをTとし，Tのうち利用者jの全Tweetを$T_j \in T$とする．通常のTweetを$h=1$，Replyを$h=2$，RTを$h=3$とし，各hの最初の投稿を$i=0$とする．そして，

表 1: 全利用者のエントロピー

	$H_{\Delta T_1}$	$H_{\Delta T_2}$	$H_{\Delta T_3}$	$H_{\Delta F_1}$	$H_{\Delta F_2}$
平均	1.977	1.815	1.180	2.556	1.892
中央値	2.018	2.005	0.278	2.673	0.667
標準偏差	1.024	1.190	1.338	1.711	2.260
最大値	4.574	4.418	4.522	11.875	11.315

表 2: 一部 Bot のエントロピー

	$H_{\Delta T_1}$	$H_{\Delta T_2}$	$H_{\Delta T_3}$	$H_{\Delta F_1}$	$H_{\Delta F_2}$
平均	1.059	0.939	0.178	2.320	0.357
中央値	0.925	0.164	0.000	1.500	0.000
標準偏差	0.861	1.175	0.618	2.615	1.268
最大値	4.185	4.278	3.855	11.875	8.355

$i-1$ 番目と i 番目の投稿の時間間隔を $\Delta t_i(h)$ とする．投稿 h の間隔 $\Delta t_i(h)$ の頻度を $n_{\Delta t_i(h)}$，最大の時間間隔を n_{T_h} としたとき，時間間隔エントロピー $H_{\Delta T_h}$ は次式で与えられる．

$$p_{\Delta T_j}\left(\Delta t_i(h)\right) = \frac{n_{\Delta t_i(h)}}{\sum_{k=1}^{n_{T_h}} n_{\Delta t_k(h)}}, \tag{1}$$

$$H_{\Delta T_h}(T_j) = -\sum_{i=1}^{n_{T_h}} p_{\Delta T_j}\left(\Delta t_i(h)\right) \log\left(p_{\Delta T_j}\left(\Delta t_i(h)\right)\right), \tag{2}$$

ここで，p は確率を表している．時間間隔エントロピーは，投稿から次の投稿までの時間差の回数を測定し，その時間差の起こりうる確率からエントロピーを算出する．

次に，Reply を $g=1$，RT を $g=2$ とし，利用者 j が利用者 i に対して行った投稿 g の全投稿 $f_i(g)$ の頻度を $n_{f_i(g)}$，利用者 j が g を行った利用者数を n_{F_g} としたとき，利用者エントロピー H_{F_g} は次式で与えられる．

$$p_{F_j}\left(f_i(g)\right) = \frac{n_{f_i(g)}}{\sum_{k=1}^{n_{F_g}} n_{f_k(g)}}, \tag{3}$$

$$H_{F_g}(T_j) = -\sum_{i=1}^{n_{F_g}} p_{F_j}\left(f_i(g)\right) \log\left(p_{F_j}\left(f_i(g)\right)\right). \tag{4}$$

4. 実験及び結果

東日本大震災前後に実際に流通した投稿をデータセットとした．2011 年 3 月 5 日-24 日の間に投稿された日本語の Tweet のうち，123 万 8612 人の利用者が投稿した 3 億 3235 万 8199 件の Tweet を分析対象とした．データは，利用者名，投稿内容，投稿時間，返信先利用者名の情報が保存されている．本研究では，1 時間単位で四捨五入して $\Delta t_i(h)$ を設定し，$H_{\Delta T}$ を算出した．

エントロピーの基本統計量を表 1 に示す．*1に日本の主な bot アカウントが 12,173 件登録されている．本研究ではここから 1000 件の bot を選出し，解析対象のデータからそれらのエントロピーを取得したものを表 2 に示す．表 1 の結果と比較すると，エントロピーの値は表 1 の結果より低いことがわかる．bot は投稿行動に対する情報量が少ないが，最大値には大きな差は見られないため，この情報量には投稿行動だけではなく別の情報も含まれていると考えられる．次に，多次元尺度構成法により，利用者を任意に抽出して可視化した結果を図 1 に示している．ここで，□を Bot，△を Cyborg[Chu 10](*1 から引用したリストのうちで，途中から地震に関する発言をしている利用者)，○を人として，それぞれ 20 人ずつ任意抽出した．ここでも Bot の違いが顕著であり，地震に関する発言を行っている Cyborg は，人間に近い特徴が示されていた．

*1 日本の Twitter Bot まとめサイト，http://bot.cuppat.net/

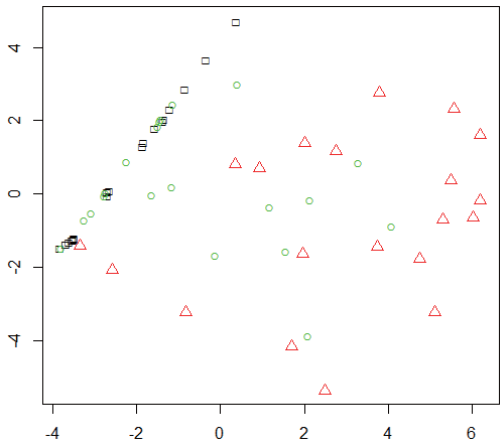

図 1: Bot，Cyborg，通常利用者の特徴の可視化

5. おわりに

本研究により，エントロピーによる利用者の特徴付けを行い，利用者の Twitter の使い方に応じた相違を確認した．本研究では，量的側面からのみの分類を行なってきたが，時間間隔パラメータの調整，Bot や Cyborg などのサンプル数を増やした場合での分析や，各エントロピー値の大小についての解釈の仕方など，様々な条件での実験試行を今後の課題とする．

参考文献

[Chu 10] Z. Chu, S. Gianvecchio, H. Wang and S. Jajodia, Who is tweeting on Twitter: human, bot, or cyborg?, Proc. of the 26th Annual Computer Security Applications Conference, pp.21-30 (2010).

[Ghosh 11] R. Ghosh, T. Surachawala, and K. Lerman, Entropy-based Classification of 'Retweeting' Activity on Twitter, Proc. of KDD workshop on Social Network Analysis (SNA-KDD), http://arxiv.org/pdf/1106.0346.pdf (2011).

[風間 12] 風間一洋, Twitter における情報伝播, 人工知能学会誌, Vol.27, No.1, pp.35-42 (2012).

[総務省消防庁 13] 総務省消防庁，大規模災害時におけるソーシャル・ネットワーキング・サービスによる緊急通報の活用可能性に関する検討会報告書，http://www.fdma.go.jp/neuter/topics/houdou/h25/2503/250327_1houdou/02_houkokusho.pdf, 2013/4/10 参照

[鳥海 13] 鳥海不二夫，大震災・そのときソーシャルメディアは動いた…のか？，第 19 回社会情報システム学シンポジウム，pp1-6 (2013).