CO-009

ゲリラ豪雨に関するツイートからのバースト検知
Bursty Events Detection from Tweets about Localized Torrential Rainfall

藤田 拓也 † 稲毛 惇人 † 大枝 真一 ‡ 中谷 剛 §
Takuya Fujita Atsuto Inage Shinichi Oeda Tsuyoshi Nakatani

1. はじめに

近年，地球の温暖化や都市部のヒートアイランド現象などによりゲリラ豪雨の被害が増加している傾向にある [1]．2008 年に日本の都市部 (東京都，神戸市，金沢市) においてゲリラ豪雨による甚大な被害が生じたのをきっかけにゲリラ豪雨という言葉は新聞やニュースなどのマスメディアで使用されるようになった．このような被害の甚大さから，ゲリラ豪雨の予測の社会的ニーズは非常に高くなっている．

防災科学技術研究所 (NIED) では，XRAIN データ [2] を活用して，ゲリラ豪雨などによる深刻な水害の早期警戒に役立てようとしている．我々は XRAIN データにゲリラ豪雨か否かラベルを与えて教師あり学習による識別器を作成を試みている．しかしながら，ゲリラ豪雨の気象学的な定義はなく，人々が遭遇する突発的な局地的大雨をゲリラ豪雨と判断することが多い [3]．つまり，XRAIN データに対して明確なゲリラ豪雨であるという教師信号を与えることができない．

そこで，我々は Twitter から雨に関する Tweet を抽出して，実際に人々がゲリラ豪雨に遭遇したかどうかを判定することで XRAIN データにゲリラ豪雨のラベルを与えることを検討してきた [4]．

しかしながら，Tweet 内容に「大雨」「豪雨」「ゲリラ」などの単語が含まれているからといって，その Tweet が本当にゲリラ豪雨に遭遇したのものであるとは限らない．そこで，本研究では雨に関する Tweet のタイムスタンプを時系列データと考え，バースト検知 [5] を行って本当のゲリラ豪雨であるかどうか判定を行う．また，バースト検知された時刻の Tweet 内容を解析して，どのような表現がなされているか調査する．

2. ゲリラ豪雨

2.1. ゲリラ豪雨の定義

「ゲリラ」は予測困難性，局地性，激甚性などの意味を持っているが [3]，ゲリラ豪雨はまさに予測が非常に困難であり，突発的で局所的な豪雨である．気象学的に明確な定義はなく，気象庁は予報用語として「ゲリラ豪雨」を用いておらず，局地的大雨などと表現している [6]．

2.2. ゲリラ豪雨の発生メカニズム

地面付近に暖かく湿った空気，上空に冷たい空気がある場合，地面付近の空気は軽く，上空の空気は重いという，大気の状態が不安定になる [6]．そこに何かのきっかけにより上昇気流が発生すると暖かく湿った空気は上空に行き，上空に行くに従って下がる気温によって，空気の水分が凝結され積乱雲となり局所的に激しい雨を降らせる．

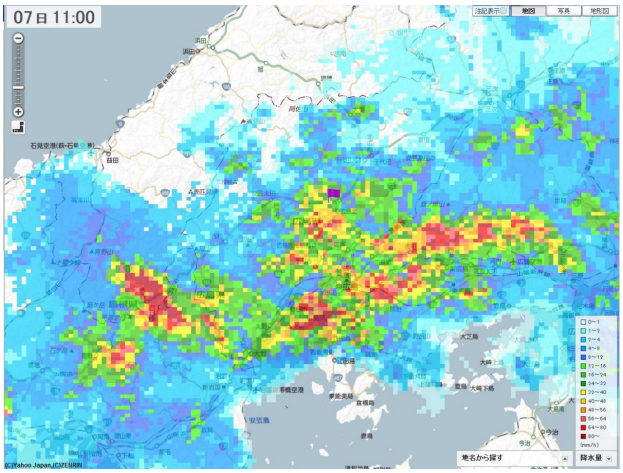

図 1: XRAIN データ [2]

2.3.XRAIN

XRAIN データの例を図 1 に示す．XRAIN は eXtended RAdar Information Network（高性能レーダ雨量計ネットワーク）の略である．また，国土交通省ではゲリラ豪雨などによる深刻な水害の早期警戒に役立てるために，平成 22 年より XRAIN によるレーダ雨量情報を提供している．従来の XRAIN は X バンド MP(マルチパラメータ) レーダ雨量計のみで構成されていたが，現在は X バンド MP レーダ雨量計と C バンド MP レーダ雨量計を組み合わけることにより，高精度・高分解能 (250m メッシュ)・高頻度 (配信間隔 1 分) のリアルタイムに近いレーダ雨量情報の配信を実現している．先行研究 [7] では，XRAIN データを用いて降雨推定を行っているものの，ゲリラ豪雨の予測を行う目的では行っていない．

3.Tweet からのバースト検知の関連研究

近年, Twitter を代表するマイクロブログの利用が盛んとなっている．Twitter は 1 回の投稿で 140 字以内という制限があり，スマートフォンやタブレットなどの携帯端末を使うことでいつでも，どこでもリアルタイムで投稿することが可能である．このような特徴から地震や豪雨などの自然災害により公共の交通機関が運行を休止した場合，その状況を投稿するユーザは少

†木更津工業高等専門学校 制御･情報システム工学専攻, Advanced Cource of Control and Information Engineering, National Institute of Technology, Kisarazu College
‡木更津工業高等専門学校 情報工学科, Department of Information and Computer Engineering, National Institute of Technology, Kisarazu College
§国立研究開発法人 防災科学技術研究所, 水・土砂防災研究部門 Storm, Flood and Landslide Research Division, National Research Institute for Earth Science and Disaster Resilience

なくない．また，Twitter の国内月間アクティブユーザ数は 4500 万人であるため，日本人の約 35%が利用しており，その普及率はさらに増加傾向にある．

Twitter ユーザが現実世界の事象に対して敏感に反応して投稿することが多いため，Twitter の投稿の傾向を分析することで，現実世界で起きた出来事や流行している話題を抽出する手法が研究されている．中でも，局所的な時間で話題の出現頻度は急激に増加する“バースト現象”の研究が注目されている．

先行研究 [8,9] では，Tweet の内容に着目し，頻出する単語のバースト性からイベントの検出を行っている．しかし，ゲリラ豪雨のような突発的な事象では，Tweet にきめ細やかな文章を記述するとは考えにくく，雨や豪雨に関する単発的なキーワードが列挙されると予想している．そこで，本研究では Tweet の内容解析ではなく，雨に関する Tweet のタイムスタンプをイベント発生の時系列データと考える．このタイムスタンプにバースト検知を行い，ゲリラ豪雨の有無について調査を行う．

4. イベント発生間隔による連続型バースト検知

時系列データにおいてイベントが急激に増加した，イベントの集中的な発生状態をバーストと呼ぶ．バーストを自動的に検出することができれば，結果としてバースト発生時の状況を効率よく解析することができるようになる．

先行研究 [5] では，2 状態の有限オートマトンを利用し，各イベントが到着した時間間隔の長さを利用して連続型バースト検知を行う．連続型バースト検知では，メッセージなどのランダム到着を表現するためによく用いられるのが指数分布である．指数分布とは，ランダムなイベントの発生間隔を表すことができる分布であり，確率密度関数が式 (1) で表されるような連続型確率分布を，平均 μ の指数分布という．x は連続型確率変数である．

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) \qquad (x \geq 0) \tag{1}$$

メッセージ到着時間間隔を生成するための最も単純なモデルは指数分布に基づく．メッセージは確率的に放出されるので，メッセージ i と $i+1$ との時間間隔 x は，パラメータ α に対する指数確率密度関数 $f(x) = \alpha e^{-\alpha x}$ となる．このモデルにおけるギャップの期待値は α^{-1} であるため，α をメッセージ到着率と呼ぶ．

4.1. 2 状態モデル

バースト検知における状態は，定常状態，バースト状態の 2 つに限定されるため，q_0, q_1 の 2 状態を持つ確率的オートマトン A を考える．A が状態 q_0 のとき，メッセージは低速で放出され，確率密度関数 $f_0(x)$ に従って独立に分布される連続メッセージ間のギャップ x をもつ．$f_0(x)$ を式 (2) に示す．

$$f_0(x) = \alpha_0 e^{-\alpha_0 x} \tag{2}$$

A が状態 q_1 にあるとき，$f_0(x)$ に従って独立に分布されるよりもギャップが短い間隔でメッセージが放出される．$f_1(x)$ を式 (3) に示す．

$$f_1(x) = \alpha_1 e^{-\alpha_1 x} \tag{3}$$

メッセージ間で A は確率 $p_q \in (0,1)$ で状態を変化させ，以前の放出および状態の変化とは無関係に，確率 $1-p_q$ で現在の状態にとどまる．

A は状態 q_0 で始まり，各メッセージ（最初のメッセージを含む）が放出される前に，A は確率 p_q で状態を変化させる．次に，メッセージが送出され，次のメッセージまでの時間間隔は，A の現在の状態に紐付いた分布に従う．

$n+1$ 個のメッセージが到着したときの，メッセージ間隔 $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ を決定してみる．ここで，メッセージ間隔 x_i は必ず正を取る．x_k に対する状態を q_{i_k} とおくことで，$\boldsymbol{x}$ に対する状態を，状態シークエンス $\boldsymbol{q} = (q_{i_1}, q_{i_2}, ..., q_{i_n})$ と表現することができ，状態シークエンス $\boldsymbol{q}$ の条件付き確率を求めるために，ベイズの定理や事後確率最大化のベイズ決定法を用いることで実現できる．

状態シークエンス $\boldsymbol{q}$ はギャップの間隔に渡って確率密度関数が式 (4) の形で導かれる．

$$f_{\boldsymbol{q}}(\boldsymbol{x}) = f_{\boldsymbol{q}}(x_1, ..., x_n) = \prod_{t=1}^{n} f_{i_t}(x_t) \tag{4}$$

もし，$q_{i_t} \neq q_{i_{t+1}}$ となるインデックスが $b = i_t$ で導かれるような，$\boldsymbol{q}$ における状態遷移数であるとき，その $\boldsymbol{q}$ の確率を式 (5) に示す．ただし，A は状態 q_0 から開始するため，$i_0 = 0$ とする．

$$\begin{aligned} p(\boldsymbol{q}) &= \left(\prod_{i_t \neq i_{t+1}} p_q \prod_{i_t = i_{t+1}} 1 - p_q \right) \\ &= p_q^b (1-p_q)^{n-b} = \left(\frac{p_q}{1-p_q} \right)^b (1-p_q)^n \end{aligned} \tag{5}$$

これから，条件付き確率 $p(\boldsymbol{q}|\boldsymbol{x})$ が求まるため，$p(\boldsymbol{q}|\boldsymbol{x})$ を式 (6) に示す．

$$\begin{aligned} p(\boldsymbol{q}|\boldsymbol{x}) &= \frac{p(\boldsymbol{q}) f_{\boldsymbol{q}}(\boldsymbol{x})}{\sum_{\boldsymbol{q}'} p(\boldsymbol{q}') f_{\boldsymbol{q}'}(\boldsymbol{x})} \\ &= \frac{1}{\sum_{\boldsymbol{q}'} p(\boldsymbol{q}') f_{\boldsymbol{q}'}(\boldsymbol{x})} \left(\frac{p_q}{1-p_q} \right)^b (1-p_q)^n \prod_{t=1}^{n} f_{i_t}(x_t) \end{aligned} \tag{6}$$

式 (6) を最大化することによって，事後確率最大化を行うことができる．また，式 (6) を最大化することは，負の自然対数をとって最小化することと等価であるため，$-\log(p(\boldsymbol{q}|\boldsymbol{x}))$ をベイズの定理を用いて式 (7)

に示す．

$$-\log\left(p\left(\boldsymbol{q}|\boldsymbol{x}\right)\right) = \log\left(\sum_{\boldsymbol{q}'} p\left(\boldsymbol{q}'\right) f_{\boldsymbol{q}'}\left(\boldsymbol{x}\right)\right) + b\log\left(\frac{1-p_q}{p_q}\right) - n\log\left(1-p_q\right) + \left(\sum_{t=1}^{n} -\log\left(f_{i_t}\left(x_t\right)\right)\right) \tag{7}$$

式 (7) の第 1 項と第 3 項は $\boldsymbol{q}$ に無関係な変数であるため，$\boldsymbol{x}$ が与えられたときの $\boldsymbol{q}$ に対するコストを求める際は無視できる．よって，式 (7) を最小化するために最小化すべきコスト $c\left(\boldsymbol{q}|\boldsymbol{x}\right)$ を式 (8) に示す．

$$c\left(\boldsymbol{q}|\boldsymbol{x}\right) = b\log\left(\frac{1-p_q}{p_q}\right) + \left(\sum_{t=1}^{n} -\log\left(f_{i_t}\left(x_t\right)\right)\right) \tag{8}$$

オートマトン A に対して，b を変化させることによって，A を現在の状態に固定する “慣性”として扱うことができるハイパーパラメータである．つまり，状態変化のしやすさを調整できる．

4.2. 無限状態モデル

期間の長さ T に渡って到着する $n+1$ 個のメッセージ間隔を考える．メッセージが T に完全に均等な間隔で到着した場合，長さ $g = \frac{T}{n}$ のギャップで到着する．高強度バーストは長さ g よりもずっと小さいギャップに近づいている．つまり，可能なバーストの全範囲を捕捉するために，任意に小さいギャップサイズに対応できるような状態を持つ無限状態オートマトンを考えるべきである．ここで，前節までに説明してきたように，基本的な目標はコストが最小の状態 $\boldsymbol{q}$ を見つける手順と同様に，コストモデルを利用する．

ここで，完全に均等な間隔で到着したときの到着率 $\alpha_0 = g^{-1} = \frac{n}{T}$ を伴い，指数分布の確率密度関数 f_0 を持つ定常状態 q_0 を有するオートマトンを考える．そのとき i $(i > 0)$ に対して，到着率 α_i を伴い f_i を持つ状態 q_i が存在する．ハイパーパラメータ s を用いて α_i を式 (9) に示す．

$$\alpha_i = \hat{g}^{-1} s^i \qquad (s > 1) \tag{9}$$

言い換えれば，状態 $q_0, q_1, \cdots$ から幾何学的に減少する到着間隔のギャップをモデル化する際に，i がより大きい値であるほどメッセージ到着の予想間隔が大きくなるように α_i が存在するということである．また，すべての i, j について，q_i から q_j への状態遷移にかかるコスト $\tau(i,j)$ が存在する．ここで，低強度バーストから高強度バーストに移行するコストについては，i, j の数値の差に比例する用に $\tau(\cdot,\cdot)$ を定義するが，高強度バーストから低強度バーストに降下するときのコストは 0 である．ハイパーパラメータ γ を用いて i, j によるコスト $\tau(i,j)$ を式 (10) に示す．

$$\tau(i,j) = \begin{cases} (j-i)\gamma & (j > i) \\ 0 & (j < i) \end{cases} \qquad (\gamma > 0) \tag{10}$$

このオートマトンは，紐づくハイパーパラメータ s および γ とともに，$A^*_{s,\gamma}$ で表される．メッセージ到着間の正のギャップ $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ が与えられるときの最小化すべきコストを式 (11) に示す．

$$c\left(\boldsymbol{q}|\boldsymbol{x}\right) = \left(\sum_{t=0}^{n-1} \tau\left(i_t, i_{t+1}\right)\right) + \left(\sum_{t=1}^{n} -\log\left(f_{i_t}\left(x_t\right)\right)\right) \tag{11}$$

状態 $q_0, q_1, ...$ に関して，$q_0, q_1, ...$ は無限に続くため，コストの最小値が明確に定義されていることを自動的に宣言することはできない．これまでのように，第 1 項を最小化することは，状態遷移数が少ないこと，およびいくつかの異なる状態にのみ遷移することと一貫しており，第 2 項を最小化することは到着間隔に近い到着率を有する状態を遷移することと一致する． どちらの最小化も本質は，状態をあまり変化させることなく可能な限りギャップの間隔を追跡することである．

また，パラメータ s, γ に関しては，s がスケーリングパラメータと呼ばれ，各状態間距離がどの程度離れているかを調整するハイパーパラメータである．γ は，より高強度バーストへの状態遷移のコストが増えるため，バースト検知の感度といえる．s は状態の離散的な間隔値が実数値のギャップを追跡することができる “解像度”であり，γ はオートマトンが状態を変えることができる容易さを制御する．

メッセージ到着間の正のギャップ $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ が与えられたとき，$A^*_{s,\gamma}$ における状態シーケンス $\boldsymbol{q} = (q_{i_1}, ..., q_{i_n})$ のコスト $c(q|x)$ を最小化する アルゴリズムの問題について考える．最小値が明確に定義され，それを計算する手段を得るためには，オートマトンのバーストレベル状態数 $q_0, q_1, ...$ が自然数 k であるように有限の制約を定義することが有用である．そして，自然数 k に対して $q_0, q_1, ..., q_{k-1}$ を $A^*_{s,\gamma}$ から求め，得られた k 状態バーストレベルオートマトンを $A^k_{s,\gamma}$ で表す．2 状態オートマトン $A^2_{s,\gamma}$ は，前述の確率的 2 状態モデルと本質的に等価であることに留意されたい．$A^*_{s,\gamma}$ における最小コスト状態シークエンス $\boldsymbol{q}$ の計算量は，有限の制約に依存するが計算は可能である．

5. 計算機実験

使用する Twitter のデータは，2013 年 6 月 1 日から 2013 年 10 月 31 日の Twitter データに対し，予め「大雨」「洪水」「豪雨」などのゲリラ豪雨に関するキーワード検索を行い，4,847,582 レコードの Twitter データを抽出している．これらの Tweet のタイムスタンプのヒストグラムを図 2 に示す．横軸は時間，縦軸は 1 分間あたりの Tweet 数を示している．図 2 を見ると，ゲリラ豪雨に関する Tweet には，メッセージ到着が均等ではなく，バースト性があることがわかる．

そこで，2013 年 7 月 23 日に着目してバースト検知を行った．図 3 にこの日の 1 分間あたりの Tweet 数を示す．この日の Tweet 数は 16 時と 17 時をピークに多くの投稿があったことがわかる．また，ハイパーパラメータ $s = 2.0, \gamma = 1.0$ としてバースト検知を行った実験結果を図 4 に示す．これを見ると，15 時から 18 時

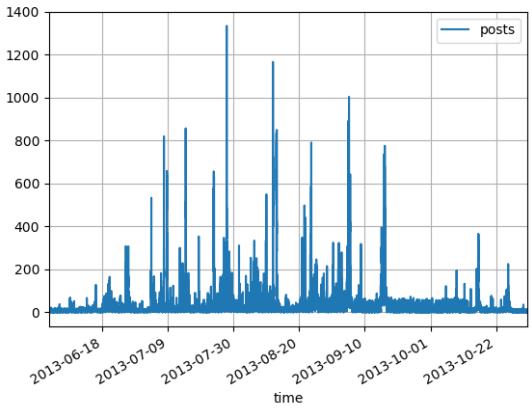

図 2: ゲリラ豪雨に関する Tweet のタイムスタンプのヒストグラム

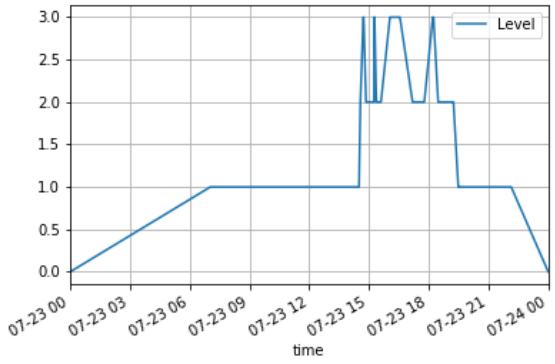

図 4: 2013 年 7 月 23 日のバースト検知 ($s = 2.0, \gamma = 1.0$)

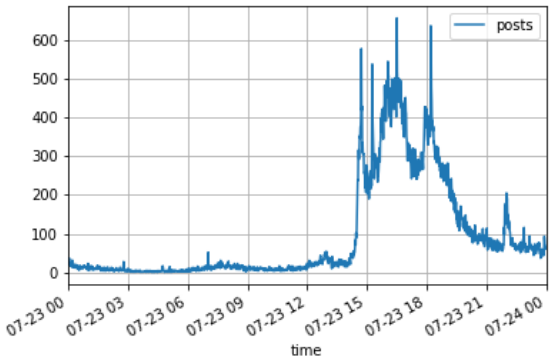

図 3: 2013 年 7 月 23 日のヒストグラム

に 4 回のバーストが検知されている．実際に，XRAIN データで確認すると 2013 年 7 月 23 日は 15 時から 17 時にかけて関東地方で局地的な大雨があったことが記録されている．また，当日のニュースとして東京都 23 区西部において雷を伴った猛烈な雨が降ったことが報道されている．この結果，雨に関する Tweet のタイムスタンプからゲリラ豪雨を検知できることがわかった．

6. まとめ

本研究では，雨に関する Tweet のタイムスタンプを時系列データと考えバースト検知を行った．バーストが起きている時刻で，本当にゲリラ豪雨が発生しているか確認するため，XRAIN データを用いて検証したところ，確かに局地的な大雨があったことがわかった．しかし，本研究では Tweet のタイムスタンプのみしか見ていないため，場所の特定はできていない．そこで，今後は，Tweet 内容に対して自然言語処理的なアプローチを行い，場所の特定を試みる．

参考文献

[1] 牛山素行, “「ゲリラ豪雨」と災害の関係について”, 水工学論文集, Vol.55, pp.S_507, 2011.

[2] XRAIN, 国土交通省, “http://www.river.go.jp/x/”.

[3] “知ってるようで知らない【ゲリラ豪雨】ってこんな雨”, “https://matome.naver.jp/odai/2147327178689566001”, 画像出典:“osero0803.blog.so-net.ne.jp”.

[4] 藤田拓也, 大枝真一, 中谷剛, “ゲリラ豪雨パターン分類のための Twitter を用いたラベル付け自動化”, 情報処理学会第 80 回全国大会, 3ZC-08, 2018.

[5] Jon Kleinberg, “Bursty and hierarchical structure in streams”, In *Proc. 8th SIGKDD*, pp.91-101, 2002.

[6] “局地的大雨から身を守るために, - 防災気象情報の活用の手引き -”, 気象庁, 2019.

[7] 加藤敦, 真木雅之, 岩波越, 三隅良平, 前坂剛, “X バンドマルチパラメータレーダ情報と気象庁レーダ情報を用いた降水ナウキャスト”, 水文・水資源学会誌, **22**(5), pp.372-385, 2009.

[8] 深沢知明, 高島真之介, 羽山徹彩, “Twitter データを用いたテレビ番組のイベント検出に関する研究”, 情報処理学会第 77 回全国大会, 3M-08, 2015.

[9] 坂本翼, 廣田雅春, 横山昌平, 福田直樹, 石川博, “Twitter ストリームのバーストの断続性に着目したキーワード抽出”, 第 4 回データ工学と情報マネジメントに関するフォーラム, C7-3, 2012.