D-013

ストリームデータからイベント収束を表すデフレーションを検出する手法の提案

Proposal of method to detect a deflation representing an event convergence from stream data

豊島 拓磨† 大野 成義† 遠藤 雅樹† 菊池 拓男†
Takuma Toyoshima Shigeyoshi Ohno Masaki Endo Takuo Kikuchi

1. はじめに

様々な物事の情報化が進んだ現代社会では、多種多様な情報をリアルタイムに取得する手段が多く存在する。その一つである Twitter は、「ツイート」と呼ばれる 140 文字以内の短文投稿を共有するマイクロブログサービスである。我が国をはじめ全世界で広く普及し、近況を気軽に投稿できるメディアとして多数のユーザに利用されている。スマートフォンを通じて容易にジオタグと呼ばれる位置情報を付与して投稿できることから、どこで何が起こっているのか即時に発信することができるソーシャルメディアである。これらの特徴から、Twitter をはじめとするソーシャルメディアはソーシャルセンサとしての利用価値が高い。物理センサを使わずして実世界を観測し、その結果を信用性の高い情報源として活かす研究が広く行われている。[1]

ジオタグ付きツイート投稿の内、特定の場所での投稿が多数ある場合、その場所での人々の集中を読み取ることができる。特定の時刻に特定の場所で投稿が集中したときなんらかのイベントや公共交通機関の異変、事故に起因する人々の停滞から不自然な集中を起こす事象を観測できる。これに対して、何らかの原因でそれまでに特定の場所に集中していたツイート数が減少した場合、停滞が解消し、たとえその場所に人々が多数集まっていたとしても絶え間なく流れる状態に変化したことを観測できる。ゆえに、テレビや新聞、ニュースサイトといった他メディアの報道よりも実世界の動向が即座に反映される。ソーシャルメディアから得られる実世界の現況は直近の予定や行動を決定する要因となる。

本研究では Twitter ユーザによって発信されるジオタグ付きツイートを大量に収集し、特定の場所、時刻におけるツイート数推移を分析する。例えばジオタグ付きツイートの収集、すなわちバースト状態を観測することでイベントやトラブルの発生をリアルタイムに検出できることを示す。さらに、リアルタイムに得られた人々の集中を表す様相を分析することにより、その後に訪れる人々の分散、あるいはイベントやトラブルの収束を表すデフレーションを検出する手法を提案する。リアルタイムに得られるデフレーションは、例えば、鉄道運行の大幅な遅延・運休をはじめとする数十万人規模の人が直近の予定について意思決定を迫られる際に必要である。単に混雑の発生している事実だけでなく、その収束をリアルタイムに抽出できるか実験・考察を行う。

本論文の構成は以下の通りである。2 章では関連研究としてバーストを検出する手法に関する研究とツイートを用いてイベント検出を行った研究について述べる。3 章では本研究における提案手法を述べる。4 章では予備実験で得られた実験結果を提示し、提案手法を適用した評価実験の結果と比較し考察する。最後に第 5 章で本研究のまとめと今後の課題について述べる。

2. 関連研究

SNS などの普及により今後もデジタルデータの量は飛躍的に増大することが予想される。これを背景に、大量のデジタルデータの有効活用に関連した研究が数多く行われている。Kleinberg[2]は時系列データのうちテキストストリームのバーストをモデリングし、構造を抽出する手法について議論している。各トピックにおけるバーストの期間、度合い、重さを表すことができるため、利用用途が広く様々な応用研究で利用されている。蝦名ら[3][4]はデータストリーム中のバーストをリアルタイムに検出する手法を提案している。監視イベントを単語の出現の有無とし、直前の状態よりも到着頻度が急激に高くなっている期間を発見することによりにバーストを検出した。特にイベントの集中発生時に保持するデータを圧縮することで計算量を抑え、リアルタイム性の高いバースト解析を実現している。一方で、マイクロブログから発信された情報を分析することにより実世界の動向を把握した上で、数十万人規模の意思決定支援に結びつける研究が様々な分野で行われている。遠藤ら[5]は生物季節観測の観光情報提示に有用な生物の見ごろを Twitter から推定する手法を提案した。日本国内で発信されるジオタグつきツイートを対象に、生物名と生物名に共起する地名や観光スポット名を基準に生物の見頃推定を行った。土屋ら[6]は鉄道運行トラブル発生時の意思決定支援を目的に Twitter を解析することによって首都圏の鉄道運行トラブルの検出および継続時間と連鎖の予測を行った。路線名を含むツイート内容を解析し、現在の運行トラブル状況について述べているツイートを手掛かりとした。その上で、鉄道運行トラブルに関連するツイートを SVM によって分類し、時系列データを生成した後、蝦名らが提案したリアルタイムバースト検出の手法を適用している。

このように、バーストをリアルタイムに検出する手法やイベントやトラブルなどがどの程度継続するのか予測する手法については多くの議論がなされているが、我々は特定の場所でリアルタイムに得られるバーストから、その収束を表すデフレーションを即時に検出する手法について提案する。蝦名らが提案したリアルタイムバースト検出の手法に遠藤らが提案した生物季節観測のための見頃推定に適用している移動平均法を組み合わせた手法である。特定の場所での人々の集中だけでなく、ツイートの減少により、通

†職業能力開発総合大学校 ,Polytechnic University

常状態へ戻る様相をリアルタイムに検出することを目的としている。

3. 提案手法

イベントの収束を表すデフレーションを検出するために、人々が特定の場所で集中および散開する現象をリアルタイムに捉える必要がある。蝦名らは監視イベントを単語の出現の有無であったのに対し、我々は個々のツイート投稿時刻とした。時刻$t-1$に投稿されたツイートの時間間隔$c(h,t)$を生成する。レベルhは時間間隔の個数を表している。図 1 に時間間隔$c(h,t)$の計算の例を示す。1 つの投稿間隔に対応するデータを得る度に、個々のツイート時刻tについて、ツイートされた時刻の間隔を求める。また、$c(h,t)$を$h+1$で割った数を$avg(c(h,t))$と表す。これはツイート間隔の平均を意味する。

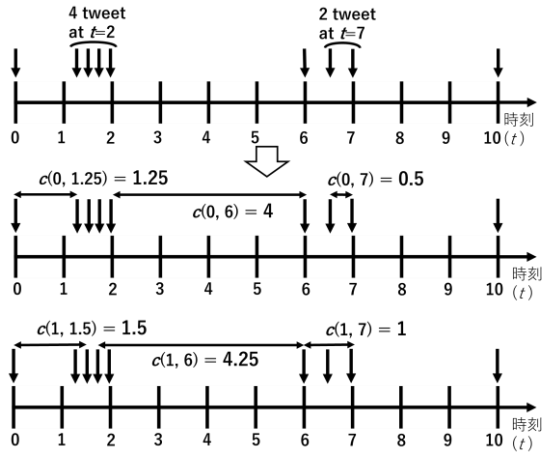

図 1 $c(h,t)$の例

ツイートの投稿時刻は秒が最小単位である。そのためツイート投稿が 1 秒間で複数あった場合が考えられる。図 1 の上段の数直線を参照すると$t=2$および$t=7$の部分がその例として挙げられる。1 秒間で 4 件と 2 件が同時に投稿されている。この場合、時間間隔$c(h,t)$の値は 0 であるか決定されない。そこで、図 1 の中段、および下段のように分割して投稿されたと仮定し、時間間隔$c(h,t)$の値を一意に決定するようにした。図 1 で表される$c(h,t)$の値を表 1 に示す。

表 1.図 1 の$c(h,t)$の値の例

h	$c(h,1.5)$	$c(h,2)$	$c(h,6)$	$c(h,6.5)$	$c(h,7)$	$c(h,10)$
0	0.25	0.25	4	0.5	0.5	3
1	1.5	0.5	4.25	4.5	1	3.5
2		0.75	4.5	4.75	5	4
3		2	4.75	5	5.25	8
4			6	5.25	5.5	8.25

蝦名らが提案した手法は、$c(h,t)$の値を比較する方法である。本研究のツイートの投稿間隔が狭くなることで投稿が増加したかの判定に適用できる。また、蝦名らが用いている種々のパラメータは時間間隔同士の比較を調整するために用いる。例として、ツイートが同一アカウントによって非常に短い間隔で投稿された場合に、わずかな変化でバースト発生と判断することを防いでいる。感度を下げることによりノイズを抑制することができる。しかし、蝦名らによる手法はバースト検出を可能としているが、バースト状態が収束したかどうかを決定することはできない。そこで我々は、遠藤らの手法を適用する。算出期間が異なる移動平均を比較し、特定の場所において人々の集中が収束するタイミングを判定する。人々の集中が発生したと判定するための条件を(1)式に定める。

$$avg\big(c(a,t)\big) < avg(c(b,t))$$
$$\text{and} \quad \beta * avg\big(c(oneday,t)\big) > avg(c(a,t)) \qquad (1)$$

また、人々の集中が収束し、散開したと判定するための条件を(2)式に定める。

$$avg\big(c(a,t)\big) > avg(c(b,t))$$
$$\text{and} \quad \beta * avg\big(c(oneday,t)\big) < avg(c(b,t)) \qquad (2)$$

ここで、$avg\big(c(oneday,t)\big)$とは、tから 24 時間前までに投稿されたすべてのツイートの時間間隔の平均値を意味する。また定数$\beta(0<\beta<1)$は、判定に用いてパラメータとして用いる。

4. 分析データ抽出および実験

4.1 分析データ抽出

イベントの収束を表すデフレーションを検出するためには、まず人々が特定の場所で集中および散開する現象をリアルタイムに捉える必要がある。本研究の提案手法と比較するために、蝦名らによる提案手法と遠藤らによる提案手法を我々が収集したデータセットに適用し、結果を提示する。本実験における対象イベントは 2019/5/4 から千葉市蘇我スポーツ運動公園で 3 日間開催された JAPAN JAM 2019 とした。JAPAN JAM は日本で開催される主要なロックミュージックイベントの 1 つである。イベント開催時間中は、開場(9:30)から最終のアーティスト終演時間(20:00)まで滞在する人や、一部のアーティストの公演時間のみ滞在する人もいることから、人々の動きの変化を見やすいと考えた。検証に用いたデータはイベント開催前日の 2019/5/3 とイベント初日 2019/5/4 に収集できたツイートのうち、ツイートに付与されたジオタグが”place_type:city”で、なおかつ千葉市中央区を表すジオタグである

[(35.5391,140.081),
(35.6177,140.081),
(35.6177,140.184),
(35.5391,140.184)]

が付与されたツイート 1845 件(2019/5/3 のツイート 597 件、2019/5/4 のツイート 1275 件)をイベント検出のデータセットとして用いた。

4.2 実験

4.2.1 比較実験の結果（蝦名らの提案手法を参考として）

比較のために、蝦名らの提案手法による実験結果を図 3 に示す。グラフの水平軸はツイートの投稿時間を表す。グラフの左側は早朝の時間帯を指しており、折れ線が左上に位置していることから、ツイート投稿間隔が長くなっていることを表している。グラフの右側は深夜帯で、ツイート

の投稿間隔に広がりが見られる。棒グラフが表す範囲は蝦名らのバースト判定結果を表している。特定のツイートとその直前に投稿されたツイートとの間隔がそれ以前より短くなっただけでバーストと判定している。蝦名らの提案手法によるバーストの判定条件式を(3)式に表す。図 3 の結果から、バーストと判定した結果が過多で、イベント会場に人々が少ない時間帯でもバーストを検出している。

$$\begin{aligned} & avg\big(c(0,t)\big) < 0.4 * avg(c(5,t-1)) \\ \text{or}\quad & avg\big(c(1,t)\big) < 0.4 * avg(c(5,t-2)) \\ \text{or}\quad & avg\big(c(2,t)\big) < 0.4 * avg(c(5,t-3)) \\ \text{or}\quad & avg\big(c(3,t)\big) < 0.4 * avg(c(5,t-4)) \\ \text{or}\quad & avg\big(c(4,t)\big) < 0.4 * avg(c(5,t-5)) \end{aligned} \tag{3}$$

我々が実験に用いたデータセットを用いて(3)式に適用した結果、バーストを過度に検出してしまい、人々の集散を表す様相を読み取ることができない。そこで、(4)式に表すように蝦名らの手法を変更し、全レベルで不等式が成り立つ条件でバースト判定を行った。またその結果を図 4 に示す。蝦名らの提案手法に比べて、バーストと判定される領域は狭まったが、早朝の時間帯に見られるバースト判定は変わらず検出されているため、他の手法を検討する必要がある。

$$\begin{aligned} & avg\big(c(0,t)\big) < 0.4 * avg(c(5,t-1)) \\ \text{and}\quad & avg\big(c(1,t)\big) < 0.4 * avg(c(5,t-2)) \\ \text{and}\quad & avg\big(c(2,t)\big) < 0.4 * avg(c(5,t-3)) \\ \text{and}\quad & avg\big(c(3,t)\big) < 0.4 * avg(c(5,t-4)) \\ \text{and}\quad & avg\big(c(4,t)\big) < 0.4 * avg(c(5,t-5)) \end{aligned} \tag{4}$$

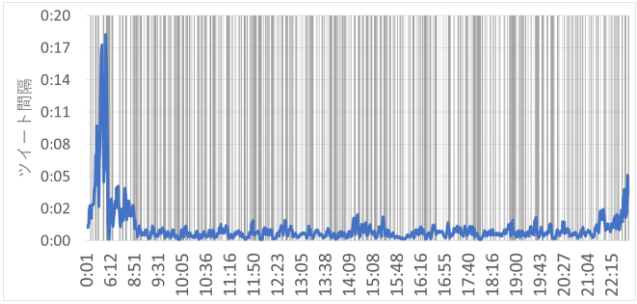

図 3.蝦名らの手法

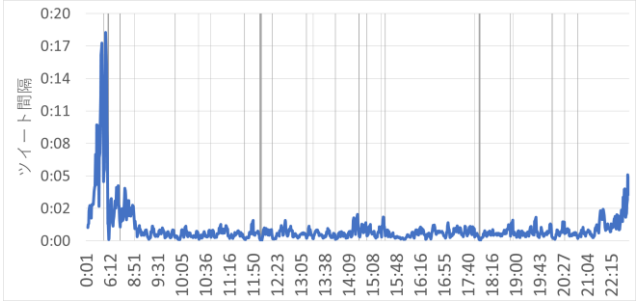

図 4.全レベルで不等式が成り立つ条件

4.2.2 比較実験結果（遠藤らの提案手法を参考として）

遠藤らは一定期間のツイート投稿頻度を分析に用いている。ツイートの投稿頻度とツイート投稿間隔は相互関係にある。ツイート投稿間隔は個々のツイート投稿時刻に依存するため、リアルタイム性が高い。一方で、ツイート頻度を計算する際には一定の時間間隔を決定しなければならない。通常、この時間間隔はツイート投稿間隔よりも大きくなるため、リアルタイム性は低くなる。遠藤らの手法では、ツイート投稿頻度から算出できる移動平均を用いて、桜の見頃期間を推定している。1 日毎の頻度分布を計算し、5 日移動平均と 7 日移動平均の大小を比較している。比較した結果によって桜の見頃を推定しており、5 日間の移動平均が 7 日間の移動平均よりも大きく、かつ前年の移動平均よりも大きくなるという条件の時、桜の見頃であると推定している。我々は、遠藤らの手法を参考に、1 日毎ではなく 5 分毎の頻度を算出し、短い時間間隔を用いてツイート間隔の変化を観察した。時刻tで生成される 5 分間隔のツイート頻度を$e(1,t)$と表す。つまり、$e(j,t)$は時刻$(t-5j)$から時刻tまでにツイートされた$5j$分間のツイート数であることを意味する。$e(j,t)$をjで割った平均値は$avg(e(j,t))$で表される。この関数 e を利用して、特定の場所において観測される人々の集中を推定する条件を以下のように設定している。

$$\begin{aligned} & avg\big(e(a,t)\big) > avg(e(b,t)) \\ \text{and}\quad & avg\big(e(a,t)\big) > avg(e(oneday,t)) \end{aligned} \tag{5}$$

人々が集中する時間帯を推定する手法に対して、人々が散開する時間帯を推定する推定手法についても検討する。そこで、遠藤らの手法を変更し、人々の集中を推定する逆条件として、人々が散開したと推定する条件を以下のように設定する。

$$\begin{aligned} & avg\big(e(a,t)\big) < avg(e(b,t)) \\ \text{and}\quad & avg\big(e(b,t)\big) < avg(e(oneday,t)) \end{aligned} \tag{6}$$

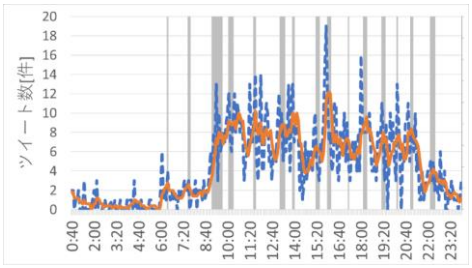

図 5.(5)式による実験結果($a=5, b=7$)
（点線:25 分移動平均　実線:35 分移動平均）

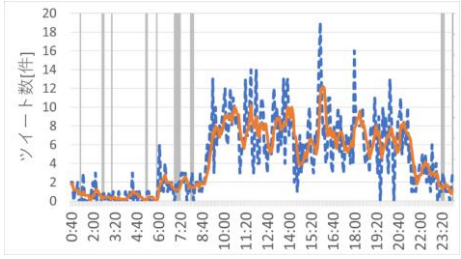

図 6.(6)式による実験結果($a=5, b=7$)
（点線:25 分移動平均　実線:35 分移動平均）

図 5 に人々が集中する時間帯の推定結果を、図 6 に人々が散開する時間帯の推定結果を示す。遠藤らの 5 日移動平均と 7 日移動平均の比較を参考に、$a = 5, b = 7$を条件とした場合、25 分間の移動平均と 35 分間の移動平均を比較していることになる。図の棒線部は推定条件を満たす時間帯を示している。$a = 5, b = 7$の条件ではツイート数の変化を敏感に推定し、早朝に人々が集中した推定結果が見られるため、時間間隔を広げた条件で実験を行った。$a = 3, b = 10$の場合の実験結果を図 8 と図 9 に示す。

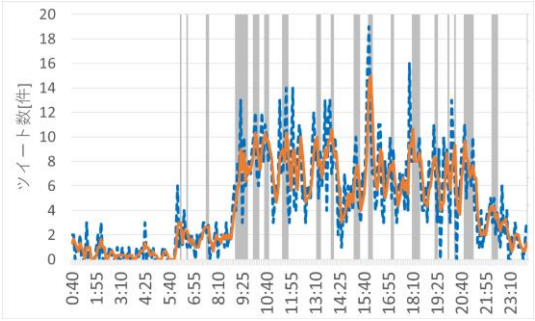

図 8.⑸式による実験結果($a = 3, b = 10$)
（点線:15 分移動平均　実線:50 分移動平均）

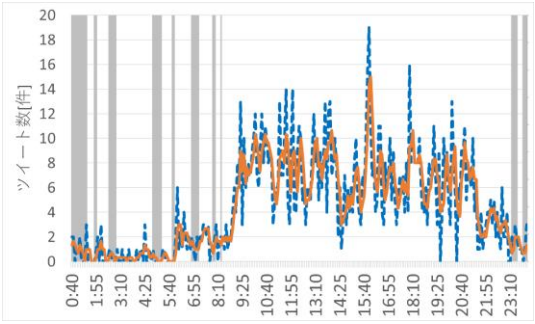

図 9.⑹式による実験結果($a = 3, b = 10$)
（点線:15 分移動平均　実線:50 分移動平均）

4.2.3 提案手法による実験

提案手法による実験結果を図 10 と図 11 に示す。リアルタイム性を実現するために、ツイート投稿頻度ではなくツイートの投稿間隔を用いている。しかし、ツイート間隔を用いた蝦名らのバースト判定手法では人々の集中を過度に検出してしまい、人々の散開を表す様相が読み取ることができない。そこで、遠藤らの手法を参考に移動平均法を用いた判定を行った。⑴式および⑵式のパラメータはそれぞれ、$a = 5, b = 30$とした。また、蝦名らの手法を参考に$\beta = 0.4$としている。図 10 の棒線部が人々の集中が発生したと判定されたことを表している。図 11 の棒線部は人々の散開を判定している。図の左側にある、人々の散開と集中の境目を表す時刻はイベントエリアが開場する直前の 9:15 頃である。一方、図の右側の境目を表す時刻は最後のアーティストのパフォーマンス終了予定時刻から 1 時間後の 21:00 前後である。閉場に伴って人々が散開したことを表している。結果として、人々が集中したことを表す判定と人々の散開を表す判定を行ったことで、イベントの開始と収束の時間に一致する結果が得られた。

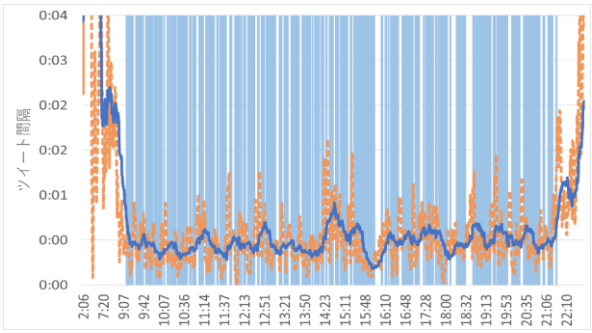

図 10.⑴式による実験結果($a = 5, b = 30$)
（点線:5 ツイート移動平均　実線:30 ツイート移動平均）

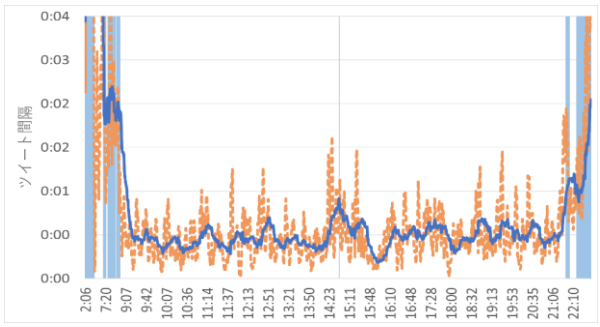

図 11.⑵式による実験結果($a = 5, b = 30$)
（点線:5 ツイート移動平均　実線:30 ツイート移動平均）

5. 本研究のまとめと今後の課題

我々はジオタグ付きツイートを分析し、リアルタイムに人々が集中する現象を捉える手法を提案した。また、人々が集中している現象だけでなく、人々が散開している状況をリアルタイムに検出することを実現したが、どのような条件でも適用できるよう確立された手法ではない。今後はパラメータを変更することによって人々の集中から散開までを捉えることを実現したい。また、⑴式と⑵式のβの値は 0.4 であり、結果としてパラメータa、bの値を最適にするために設定された値である。パラメータa、b、βの最適な値を決定するためのアルゴリズムを検討する。

参考文献

[1] 榊 剛史, 松尾 豊, “ソーシャルセンサとしての Twitter:ソーシャルセンサは物理センサを凌駕するか(<特集>Twitter とソーシャルメディア)”, 人工知能学会誌, Vol.27, No.1, pp.67-74 (2012).
[2] J. Kleinberg, “Bursty Hierarchical Structure in Streams”, In Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1-25(2002).
[3] 蝦名 亮平, 中村 健二, 小柳 滋, “リアルタイムバースト検出手法の提案”, 日本データベース学会論文誌, Vol.9, No.2 (2010).
[4] 蝦名 亮平, 中村 健司, 小柳 滋, “リアルタイムバースト解析手法の提案”, 情報処理学会論文誌, Vol.5, No.3, pp.86-96 (2012)
[5] 遠藤 雅樹, 三富 恵佑, 佐伯 恵佑, 江原 遥, 廣田 雅春, 大野 成義, 石川 博, “ツイートを用いた生物季節観測の見頃推定手法による情報提供の検討”, 観光と情報, 第 12 巻, 第 1 号 (2016).
[6] 土屋 圭, 豊田 正史, 喜連川 優, “マイクロブログを用いた運行トラブル発生期間および付帯情報の抽出”,第 6 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2014), B3-2,(2014).