F-013

双対 SR 法による Twitter データの時空間バースト検出

Spatio-temporal Burst Detection from Twitter Data Using Dual-graph SR Method

加藤翔子 † 斉藤和巳 † 風間一洋 ‡
Shoko Kato Kazumi Saito Kazuhiro Kazama

1. はじめに

Twitter，Facebook，LinkedIn などのサービスの普及により，オンラインサービス上に構築したソーシャルネットワークを経由して，多種多様な情報を効率良く送受信できるようになった．このような情報交換は，1) 各ユーザが非同期に行う個人的なつぶやき，2) 現実世界や仮想世界のイベントと連動して生じる一過性の盛り上がり，3) 多数のユーザ間のコミュニケーションにより生まれる**議論話題** の 3 種類に大きく分類できる．

Twitter について考えると，1) は大部分が個人的な内容であり，他者への影響は小さい．2) は TV 番組のクライマックスや地震などのイベント時に多くのユーザが同期して発言したり，あるユーザの発言を拡散してフォロワに転送するもので，瞬間的に大量のツイートを誘発するが，イベント発生以外の情報はほとんど述べられないことが多い．さらに，情報を充分精査せずに転送するので，デマ情報も多く含まれる．これに対し，3) は，主に全体の約 2〜4 割を占めるリプライで交換・伝播される．リプライは発言元と発言先のフォロワも閲覧できることから，フォローネットワーク上の 2 次近傍ユーザも含む緩やかな情報拡散を引き起こし，インターネット上の合意形成や口コミの拡散などの重要な社会現象に関係する．このような形式での情報伝搬はマスメディアには存在し得ないため，情報学に留まらず，経営学，社会科学，さらには心理学などの学術分野においても重要な研究対象である．

多くのユーザ間で大規模に展開される議論話題の検出，その内容の分析，さらに影響拡散のモデリングに関する技術確立への第一歩として，本稿では，コミュニケーションに時系列情報が付与されたデータで構築した双対グラフのコアを抽出する双対 SR 法を提案する．双対 SR 法のコア抽出は，ネットワークの中から結合が密なノード群のコア部を直接探索し，ノード群の重複を許容してコア部を抽出する MDSR 法 [3] を応用したものである．評価実験には東日本大震災前後における Twitter のリプライデータ [2] を用いる．

本稿の範囲では，双対 SR 法が同一のユーザから発せられた異なる話題を正しく分けて抽出することや，後述する時間の閾値の設定により，優先的に抽出する話題やユーザに差異が生じることを示す．

なお，本稿で作成するグラフは，正確な意味での双対グラフとは異なる性質を持つが，従来の分析ではリンクと考えられてきたコミュニケーションを本稿ではノードとして捉えているため，便宜的に双対という表現を用いる．

†静岡県立大学
‡和歌山大学

2. 双対 SR 法

以下では，ユーザ間の時刻付きメッセージをノードとして作成したグラフのコア抽出を行う双対 SR 法について説明する．

2.1. 双対グラフの作成

ユーザ集合 $U = \{u, v, \ldots\}$ について，ユーザ $u \in U$ からユーザ $v \in U$ へ時刻 t に送られたメッセージを (u, v, t) と表すと，全メッセージ集合は $D = \{(u_1, v_1, t_1), (u_2, v_2, t_2), \ldots, (u_I, v_I, t_I)\}$ と表記できる．以下の分析では，この D の各要素をノードとする．

いま，各メッセージの間に，ユーザ v を受信者とするメッセージからユーザ v を送信者とするメッセージへのリンクが存在すると考え，受信した情報をさらに送信したと判断するための時間の閾値 Δt を設けることで，リンク集合 $E((u, v, t))$ を以下のように定義する．

$$E((u, v, t)) = \{((u, v, t), (v, w, t')) | t < t' \leq t + \Delta t\} \quad (1)$$

なお，このときの v を特に**中心ユーザ**と呼ぶこととする．この $E((u, v, t))$ を D 内の全てのメッセージについて考えることで，全リンク集合 $E = \bigcup_{(u,v,t) \subset D} E((u, v, t))$ が得られる．

本研究で着目している議論話題は，短い時間間隔内でのコミュニケーションから生まれると推測できるため，あるメッセージを受け取ってから別のメッセージを送信するまでに掛かった時間が充分短いことを保証できるように，閾値 Δt を設定する．

このようにして得られる双対グラフ $G_{\Delta t} = (D, E)$ について，以下に示すコア抽出手法を適用することで，時空間バーストを検出する．

2.2.MDSR 法の応用

MDSR 法 [3] では，与えられたグラフの多重有向隣接行列 $\mathbf{A}$ の右固有ベクトルと左固有ベクトルを 2 値に量子化してコア部を抽出する．さらに，隣接行列から抽出したコア部に含まれるリンクを削除した後に上記の処理を適用し，再帰的にコア部を抽出する．

D の 2 つの部分集合 $W \subset D$ と $X \subset D$ に対し，その間に張られたリンク密度 $G(W, X)$ を次式で定義する．

$$G(W, X) = \frac{1}{\sqrt{|W||X|}} \sum_{i \in W} \sum_{j \in X} A(i, j). \quad (2)$$

$|W|$ や $|X|$ は，集合 W や X の要素数を表し，$G(W, X)$ はこれらの幾何平均である．この式 (2) を最大にするノード部分集合ペア W と X の探索問題を考えることで，リンクが密集しているコア部を抽出する．抽出された全ての $i \in W$ と $j \in X$ について，$A(i, j) = 0$ とし，再度 (2) 式を最大化するノード部分集合ペアを求

めることで，再帰的にコアを抽出する．以下では，k 回目の試行で抽出されたコアを C_k と表記する．

2.1 節で作成した双対グラフは単純有向グラフであるが，同様に (2) 式を最大化することでコア部を抽出できる．このときコアとして抽出される密なコミュニケーションについて内容や中心ユーザを調査する．全てのコミュニケーションの内容を調査する必要が無いため，効率的な時空間バースト検出が期待できる．

3. 実験

3.1. データ概要

本稿では，2011 年 3 月 5 日 00:00:00 から同月 24 日 23:59:59 までの日本語で投稿されたツイート [2] の中から，文頭が “@user” から始まるツイートを収集した．各ユーザから@user で指定されたユーザへの時刻付きメッセージをノードと考え，データセットとした．

3.2. 分析結果

本稿では，$\Delta t = 600, 1800, 3600, 7200$ の 4 つの場合を調査する．すなわち $G_{600}, G_{1800}, G_{3600}, G_{7200}$ を 2.1 節で説明した手法で作成し，2.2 節で説明した手法により時空間コミュニティ検出を行う．

各条件で抽出した上位 10 個のコアの中心ユーザを，表 1 に示す．各中心ユーザの属性を調査した結果，2 パターンに大別できたため，それに則してアカウント名にマスキングを施し ID をつけた．一つは bot アカウント群であり，b_1 のように表記する．もう一方は，芸能人や企業の公式ユーザのような広く知られているアカウントであり，f_1 のように表記する．以下では有名アカウントと呼ぶ．

表 1 より，双対 SR 法は，Δt を小さく設定すると bot アカウントを中心ユーザとして検出するが，Δt の値を大きくするにつれ，有名アカウントを中心ユーザとして検出する傾向にある．これは，プログラムは短い時間で機械的に応答しているが，人間同士のコミュニケーションでは，メッセージの受信から送信までにある程度の時間が必要であることを示唆している．

また，双対グラフごとに抽出されたコアを観察すると，G_{600} の C_3 以外のコアは全て，b_1 が中心ユーザとなっている．同様に，G_{1800} の C_4, C_7, C_{10} や C_6, C_8，G_{3600} の C_3, C_6，G_{7200} の C_2, C_3, C_7 や C_6, C_8 は，他のコアと重複する中心ユーザを持つ．これらのコアについてメッセージの内容を調査したところ，bot アカウントを中心ユーザとする重複コアの間には明確な話題の差異が見られなかったが，有名アカウントを中心ユーザとする重複コアの間では，明らかに違った話題についてのリプライが交わされていることを観察した．bot アカウントは決められたメッセージを機械的に発信するため，異なる時系列でも同様の話題となるが，有名アカウントは人間の手によってメッセージが発信されるため，その内容は時間とともに変化する．このような同一のアカウントから発せられた異なる話題を，双対 SR 法は正しく分けて抽出すると言える．メッセージの具体的な内容として，有名アカウントが中心ユーザとなるコアでは，芸能人の誕生日を祝う内容や，東日本大震災を受けての行動についての意見などが見られた．

表 1: 双対 SR 法で抽出したコアの中心ユーザ

C_k	G_{600}	G_{1800}	G_{3600}	G_{7200}
C_1	b_1	b_2	f_1	f_1
C_2	b_1	f_2	f_3	f_4
C_3	f_5	b_3	f_4	f_4
C_4	b_1	b_1	b_2	f_6
C_5	b_1	f_5	f_7	f_3
C_6	b_1	b_4, b_5, b_6, f_3	f_4	f_7
C_7	b_1	b_1	f_5	f_4
C_8	b_1	b_4, b_5, b_6	b_7	f_7
C_9	b_1	f_8	f_8	f_9
C_{10}	b_1	b_1	f_6	f_{10}

4. おわりに

時系列情報を伴うユーザ間のコミュニケーションのバースト検出を行う手法として，閾値を設けた双対グラフを生成し，MDSR 法の応用によりコア抽出を行う双対 SR 法を提案した．東日本大震災前後における Twitter のリプライをデータセットとした評価実験により，同一の有名アカウントから発せられた異なる話題を，双対 SR 法は正しく分けて抽出することを確認した．

また，双対グラフを生成する際の閾値の設定によって，優先的に抽出される中心ユーザに変化が生じることも確認した．今後の課題としては，他のコミュニケーションをデータセットとした評価実験により，この閾値を決定するための統計的指標の確立や更なる手法の確立，さらに，議論話題における影響拡散のモデリングなどに着手する．

また，今回はコア抽出を各 10 回試行し，bot や有名人が中心ユーザとなるコアを観察したが，試行回数を増やしより多くのコアを抽出することで，一般ユーザ同士の議論構造や，中心ユーザが小規模でなく存在するような議論構造などの発見につながる可能性があるため，引き続き調査していく．

謝辞

本研究は，科研費 (C)(No.26330345) の助成を受けた．

参考文献

[1] Chakrabarti Deepayan, Punera Kunal, ”Event summarization using tweets,” Proc. Fifth International AAAI Conference on Weblogs and Social Media, pp.66–73, 2011.

[2] 鳥海不二夫, 篠田孝祐, 栗原聡, 榊剛史, 風間一洋, 野田五十樹, ”震災がもたらしたソーシャルメディアの変化,” ネットワークが創発する知能研究会, 2011.

[3] 加藤翔子, 斉藤和巳, 風間一洋, 佐藤哲司, ”MDSR 法を用いた reply ツイートネットワークの特性分析,” WebDB Forum 2013, 2013.