DBSJ Journal, Vol.11, No.2
October 2012

情報拡散モデルに基づくツィート系列からのバースト期間検出

Burst Detection in a Sequence of Tweets based on Information Diffusion Model

大原 剛三♥　斉藤 和巳♦
木村 昌弘♠　元田 浩♣

Kouzou OHARA　Kazumi SAITO
Masahiro KIMURA　Hiroshi MOTODA

本論文では，ソーシャルネットワーク上で観測された情報拡散系列からバースト的な拡散が起こった区間を検出する手法を提案し，Twitter 上で実際に観測した情報拡散系列に適用した結果について報告する．提案法は，情報が伝搬する際に生じる遅延時間が指数分布などの確率分布に従う情報拡散モデルを仮定し，時系列上でその分布が変化する変化点を観測データから推定する事でバースト期間を検出する．実際の情報拡散データを用いた評価実験では，提案法が効率的，かつ精度よく変化点を検出できることを示すとともに，複数の話題が混在する情報拡散系列から検出したバースト期間においてもその先頭付近においては同一の話題が集中的に現れる傾向にあること，およびネットワーク上での実際の情報拡散経路が不明な場合であっても直線状の伝搬経路を仮定する事が変化点検出においてよい近似解を与え得ることを示す．

We propose a method of detecting the period in which a burst of information diffusion took place from an observed diffusion sequence data over a social network. We assume a generic information diffusion model in which time delay associated with the diffusion follows the exponential distribution and the burst is directly reflected to the changes in the time delay parameter of the distribution. The proposed method detects the change points by maximizing the likelihood of generating the observed diffusion sequence. Through experimental evaluations with the real Twitter data, we show that the proposed method can detect the change points efficiently and accurately, as well as present two interesting discoveries that the beginning of a detected burst period tends to contain massive homogeneous tweets on a specific topic even if the observed diffusion sequence consists of heterogeneous tweets, and that assuming the information diffusion path is a line shape tree can give a good approximation of the maximum likelihood estimator.

1. はじめに

近年，高性能な携帯電話，スマートフォンの普及によりTwitterやFacebookなどのソーシャルメディアを介したコミュニケーションが我々の日常生活に浸透し，大きな影響を与えつつある[1-3]．これらのソーシャルメディアにおいては，通常，友人関係などに基づいて形成されたソーシャルネットワークを介して情報が伝搬される．特に，140文字までの短文を発信するマイクロブログと呼ばれるTwitterは，その情報発信の手軽さから若者を中心に爆発的に普及しており，先の東日本大震災を含む国内外の自然災害発生時においては情報インフラとして重要な役割を果たしたことは記憶に新しい．

そのようなソーシャルネットワークに関しては，ネットワーク構造に基づきノードを特徴づける中心性指標と呼ばれる指標が幾つか提案されている[4-6]．中心性指標はソーシャルネットワーク上の情報拡散において重要な役割を果たすノードの特定に利用できるが，可能な限り多くのノードに情報を伝搬させることのできる影響度の高いノード集合を見つける影響最大化問題[7,8]などでは，情報拡散メカニズムを陽に用いることが重要となる[8]．一般には，そのメカニズムは確率モデルによってモデル化される．最も代表的で基本的なものは独立カスケード（IC: Independent Cascade）モデル[7,9]と線形閾値（LT: linear threshold）モデル[10,11]であり，我々はさらに非同期時間遅れを考慮したそれらの拡張であるAsIC（Asynchronous time delay IC）モデル[12]やAsLT（Asynchronous time delay LT）[13]モデルを提案している．実際，これらのモデルを用いて同定した影響度の高いノードやリンクは中心性指標により同定されるものとは大幅に異なる．

一方，実際の情報拡散においては，ある特定の情報が短期間に爆発的に拡散することがある．そのような情報のバースト的な拡散が何を契機として生じたのかを知ることは実世界において観測された事象を理解・分析する上で重要であり，そのためには観測データからバースト期間を精度よく特定できる必要がある．本研究では，そのようなバースト的な情報拡散現象が情報拡散モデルのもつパラメータの変化により生じたものと捉え，その変化点を観測データから検出する問題を考える．実際には，AsICやAsLTなどの時間遅れを考慮したモデルでは，その情報伝搬の遅れは指数分布などによりモデル化されるため，本研究ではその分布を規定するパラメータ（時間遅れパラメータ）にのみ着目する．これは，実際のバースト期間では特定の情報の伝搬間隔が短くなる，つまり情報伝搬の遅延時間が短くなるためであり，これにより特定のモデルを仮定することなく変化点を検出する．具体的には，時系列上の複数の変化点により生じる時間遅れパラメータの系列を観測した情報拡散系列に対して尤度比検定を用いて効率的に最適化する手法を提案する．そして，東日本大震災前後にTwitter上で発信された実際のツィートから抽出した時系列データに提案手法を適用し，効率的，かつ高精度に変化点を検出できることを実験的に示す．さらにその分析結果を通して，対象とした情報拡散系列が指定したユーザIDを含むツィートを集めた系列であり，複数の話題を含んでいても，検出した変化点に基づき特定したバースト期間の先

♥ 非会員　青山学院大学理工学部　ohara@it.aoyama.ac.jp
♦ 正会員　静岡県立大学経営情報学部
k-saito@u-shizuoka-ken.ac.jp
♠ 非会員　龍谷大学理工学部電子情報学科
kimura@rins.ryukoku.ac.jp
♣ 非会員　大阪大学産業科学研究所
motoda@ar.sanken.osaka-u.ac.jp

頭部分には同一内容のツィートが集中的に現れること，およびネットワーク上の実際の情報拡散経路が分からなくても，直線状の伝搬経路を仮定することが最尤推定値のよい近似を与え得ることを示す．

2. 情報拡散モデルの枠組み

有向グラフ $G=(V,E)$ として構造が規定されるソーシャルネットワーク上での情報拡散を考える．ここで，V と E $(\subseteq V\times V)$ はそれぞれ全ノードの集合と全リンクの集合を表す．いま，ノード v_0 からの時刻 t_0 での情報発信で起きた情報拡散系列 $C=\{(v_0,t_0),(v_1,t_1),\cdots,(v_N,t_N)\}$ を観測したとする．ここで，v_n は情報が伝播したノードで，t_n はその時刻を表すとする．また，任意の $n\in\{1,\cdots,N\}$ に対して，$t_{n-1}<t_n$ となるよう番号付けされているとする．一方，標準的な設定として，情報拡散系列 C は有向グラフ G が内包するツリーとして規定されるとする[14]．すなわち，ノード v_n に情報を伝播した親ノードは $v_{p(n)}$ として一意に確定されるとする．ここで，$p(n)$ はノード v_n の親ノード番号を $\{0,\cdots,n-1\}$ の範囲で返す関数を表す．

情報拡散系列 C を生成する基本モデルとしては，時間遅れを考慮していなかった従来のICモデル[7,9]やLTモデル[10,11]に対し，実世界での情報伝搬に見られる非同期時間遅れの概念を導入したAsICモデル[12]やAsLTモデル[13]などであるとする．すなわち，ノード v_n に情報伝播した時刻 t_n として任意の実数を許容し，情報伝播の遅延時間 $t_n-t_{p(n)}$ については，ある確率分布に従うとする．なお，本稿では議論の簡単化のために遅延時間は指数分布に従うものとするが，べき乗分布など他の分布も同様に用いることが可能である．

3. 問題設定

本節では，本稿が取り扱う“変化点検出問題”を形式的に定義する．本研究では，情報拡散過程で何らかの変化が生じ，その変化が内包された拡散系列を観察するものと仮定する．このとき，変化点検出問題では，それぞれの変化点において，その変化がいつ生じ，どのくらいの長さ持続したのかを検出することを目的とする．ここで，同じトピックを話題にする際の人々の行動は極めて類似したものになると仮定できること[12,13]から，本稿では，リンク $(u,v)\in E$ 毎に設定可能な時間遅れパラメータ $r_{u,v}$ はリンク依存せず一定の値を取るという制約をおく．すなわち，$r_{u,v}=r$ $(\forall(u,v)\in E)$ とし，情報拡散遅れは $p(t_n-t_{p(n)};r)=r\exp(-r(t_n-t_{p(n)}))$ という指数分布に従うとする．

次に，変化点検出問題を数学的に定式化する．情報拡散系列の時刻集合 $D=\{t_0,t_1,\cdots,t_N\}$ が観測されたとし，あるトピックの拡散過程における変化点を T_j とする．ここで，$t_0<T_j<t_N$ である．このとき，変化点 T_j の直前までは r_j，そして変化点 T_j 直後では r_{j+1} というパラメータの指数分布に従うとする．いま，J 個の時刻から構成される変化点集合を $S_J=\{T_1,\cdots,T_J\}$ とし，便宜上 $T_0=t_0$ かつ $T_{J+1}=t_N$ と設定しておく．また，$T_{j-1}<T_j$ であるとし，S_J による D の分割を $D_j=\{t_n;T_{j-1}<t_n\le T_j\}$ で定義する．すなわち，$D=\{t_o\}\cup D_1\cup\cdots\cup D_{J+1}$ となり，$|D_j|$ は区間 $(T_{j-1},T_j]$ に含まれる観測時刻数を表す．ここで，任意の $j\in\{1,\cdots,J+1\}$ に対して，$|D_j|\neq 0$ とし，少なくとも一つの観測時刻 t_n が存在し $t_n\in D_j$ を満たすと仮定する．一方，パラメータのベクトルを $\mathbf{r}_{J+1}=(r_1,\cdots,r_{J+1})$ で定義すれば，変化点集合 S_J が与えられたときの観測データ D に対する対数尤度は次式で計算できる．

$$\begin{aligned}&L(D;\mathbf{r}_{J+1},S_J)\\&=\log\prod_{j=1}^{J+1}\prod_{t_n\in D_j}r_j\exp(-r_j(t_n-t_{p(n)}))\\&=\sum_{j=1}^{J+1}|D_j|\log r_j-\sum_{j=1}^{J+1}r_j\sum_{t_n\in D_j}(t_n-t_{p(n)}).\end{aligned}\tag{1}$$

よって，式(1)の尤度を最大にするパラメータの最尤推定値は次のように計算できる．

$$\hat{r}_j^{-1}=\frac{1}{|D_j|}\sum_{t_n\in D_j}(t_n-t_{p(n)}),\quad j=1,\cdots,J+1.\tag{2}$$

さらに，式(2)の最尤推定値を式(1) に代入すれば以下となる．

$$\begin{aligned}&L(D;\hat{\mathbf{r}}_{J+1},S_J)\\&=-N-\sum_{j=1}^{J+1}|D_j|\log\left(\frac{1}{|D_j|}\sum_{t_n\in D_j}(t_n-t_{p(n)})\right).\end{aligned}\tag{3}$$

したがって，我々の変化点検出問題は，式(3)を最大化する変化点集合 S_J を求める問題となる．ただし，式(3)では，変化点集合 S_J 導入の効果を直接評価できない．よって，この問題の別表現として，尤度比検定の目的関数として我々の変化点検出問題を定式化する．まず，変化点が存在しないとした $S_0=\phi$ に対して，式(3)は次のように計算できる．

$$L(D;\hat{r}_1,S_0)=-N-N\log\left(\frac{1}{N}\sum_{n=1}^{N}(t_n-t_{p(n)})\right).\tag{4}$$

よって，変化点が J 個存在するとしたときと，存在しないとしたときの尤度比の対数は次式で定義できる．

$$\begin{aligned}LR(S_J)&=L(D;\hat{\mathbf{r}}_{J+1},S_J)-L(D;\hat{r}_1,S_0)\\&=N\log\left(\frac{1}{N}\sum_{n=1}^{N}(t_n-t_{p(n)})\right)\\&\quad-\sum_{j=1}^{J+1}|D_j|\log\left(\frac{1}{|D_j|}\sum_{t_n\in D_j}(t_n-t_{p(n)})\right).\end{aligned}\tag{5}$$

すなわち本論文では，式(5)で定義した $LR(S_J)$ を最大化する変化点集合 S_J を求める問題を考える．

一般に，情報拡散系列のツリー構造を完全に得るのは困難な場合も想定できる．ここでは，最も単純な両極端のケースを考える．すなわち，最も拡散時刻が早くなるスター状ツリーと，最も遅くなる一次元の直線状ツリーである．情報拡散の親ノードを規定する関数については，前者では $p(n)=0$ であり，後者では $p(n)=n-1$ となる．変化点を含まないケースでは，パラメータ推定値は，スター状ツリーでは $r^{-1}=(t_1+\cdots+t_N)/N-t_0$ となり，直線状ツリーでは $r^{-1}=(t_N-t_0)/N$ となる．現実には，これらのパラメータ

推定値の中間的な値になると想定されるが，ここでは実際のツリー構造が不明な場合にこれらいずれかの値で近似することを考える．ここで，前者の値がノードv_0から各ノードまでの平均到達時間を表すのに対し，後者の値が隣接する観測時刻間の平均間隔を表すこと，バースト期間では観測間隔が他の期間よりずっと短くなることを考慮すると，近似としては後者の直線状ツリーがより適していると考えられる．このとき，式(5)で定義した$LR(S_J)$は次のように計算できる．

$$LR(S_J) = N \log\left(\frac{t_N - t_0}{N}\right) - \sum_{j=1}^{J+1} |D_j| \log\left(\frac{T_j - T_{j-1}}{|D_j|}\right). \quad (6)$$

5.2 節では式(6)を用いた実験結果を主に示すとともに，スター状ツリーを仮定した場合の結果との比較も報告する．

4. 変化点検出法

情報拡散結果の時刻集合$D = \{t_0, t_1, \cdots, t_N\}$と変化点の個数$J$が与えられたとき，変化点集合$S_J = \{T_1, \cdots, T_J\}$を時刻集合の部分集合$S_J \subset D$として求める変化点検出法について考える．以下では，ナイーブ法，シンプル法，及び，提案法のそれぞれについて説明する．

4.1 ナイーブ法

最も単純な方法は，観測時刻集合Dに対し，虱潰しで最適なJ個の変化点集合S_Jを求めるナイーブ法である．明らかに，ナイーブ法の計算量は$O(N^J)$である．したがって，ある程度Nが大きくなると，実用的な時間で結果が得られるのは$J = 2$程度までに限定される．

4.2 シンプル法

変化点の個数Jが大きい場合にも適用可能なシンプル法について述べる．この方法では，既に選定した$(j-1)$個の変化点集合S_{j-1}を固定し，新たに付加するとして最適な変化点T_jを求めS_{j-1}に加えることを，$j = 1$からJまで繰り返す．すなわち，そのアルゴリズムは以下となる．

S1-1. $j = 1$，$S_0 = \phi$と初期化する．
S1-2. $T_j = \arg\max_{t_n \in D}\{LR(S_{j-1} \cup \{t_n\})\}$を求める．
S1-3. $S_j = S_{j-1} \cup \{T_j\}$と更新する．
S1-4. $j = J$ならS_Jを出力し終了．
S1-5. $j = j + 1$とし，S1-2 へ戻る．

ただし，S1-3 において，変化点集合S_jの要素は$T_{i-1} < T_i$となるようインデックスを更新するとする．ここで，$i = 2, \cdots, j$である．明らかに，シンプル法の計算量は$O(NJ)$となる．すなわち，ある程度Nが大きくなっても，任意のJに対して，実用的な時間で結果を得ることができる．しかしながら，シンプル法は貪欲法に基づく手法であるため，比較的プア―な局所解にトラップされるケースも危惧される．

4.3 提案法

前述のシンプル法と同程度の計算量で，解品質の向上を目的とした提案法について述べる．この方法では，シンプル法で求めた変化点集合S_Jを出発点として，既に選定した一つの変化点T_jを選び，これ以外の変化点集合$S_J \setminus \{T_j\}$を固定し，より望ましい別の変化点T'_jに置き換えることを繰り返す．ここで，$\cdot\setminus\cdot$は集合差を表す．明らかに，$j = 1, \cdots, J$のすべてで置き換えできなければ，すなわち，どのjでも$T'_j = T_j$となれば，この方法でさらなる改善はできないので反復を終了させるとする．すなわち，そのアルゴリズムは以下となる．

S2-1. S_Jをシンプル法で求め，$j = 1$，$k = 0$と初期化する．
S2-2. $T'_j = \arg\max_{t_n \in D}\{LR(S_J \setminus \{T_j\} \cup \{t_n\})\}$を求める．
S2-3. $T'_j = T_j$なら$k = k + 1$，さもなければ$k = 0$とし，$S_J = S_J \setminus \{T_j\} \cup \{T'_j\}$と更新する．
S2-4. $k = J$ならS_Jを出力し終了．
S2-5. $j = J$なら$j = 1$，さもなければ$j = j + 1$とし，S2-2 へ戻る．

明らかに，シンプル法と比較して，一般に，提案法は数倍の計算量が必要となる．ただし，計算量がどの程度増加するかとともに，解品質がどの程度改善するかは問題に依存する．本稿では，これらについて実データを用いた計算機実験により評価する．

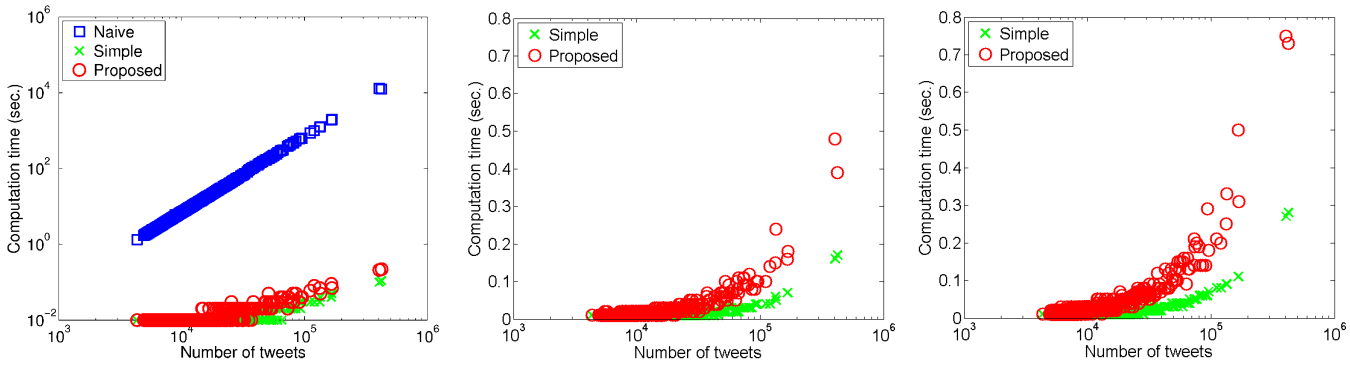
5. 評価実験

5.1 実験設定

本稿では，Twitter 上の実際の情報拡散系列を用いて，前節で述べた各変化点検出法の計算時間，および変化点検出精度を実験的に評価した．本実験で用いた情報拡散系列は，東日本大震災が発生した 2011 年 3 月 11 日を含む 2011 年 3 月 5 日から 3 月 24 日までの約 3 週間に 200 件以上のツィートをした 1,088,040 名の Twitte ユーザのツィート 201,297,161 件からから抽出したものである．本実験では，他のユーザのツィートをそのまま発信するリツィートを情報の拡散と捉え，収集した全ツィートから各ユーザの ID を“@ID”形式で含むツィートを抽出し，そのユーザごとの時系列データを情報拡散系列として用いた．これは，厳密にリツィートの系列のみを抽出することが困難な為である．具体的には，系列長（ツィート数）が 5,000 以上の 798 名のユーザに対する情報拡散系列を用いた．ここで，各系列にはリツィート以外のツィート，および複数の話題に関するリツィート系列が含まれ得ることに注意されたい．これらの情報拡散系列に対して，厳密解を与えるナイーブ法を現実的な時間内で実行するために本実験では変化点数の基準値を$J = 2$とした．また，実際の情報拡散経路は不明である為，前述のように直線状ツリーを仮定し，式(6)を用いて尤度を計算した．なお，次節で述べる実験結果はいずれも Intel(R) Xeon(R) CPU W5590 @3.33GHz，メモリ 32GB をもつ計算機上で実行したものである．

5.2 実験結果

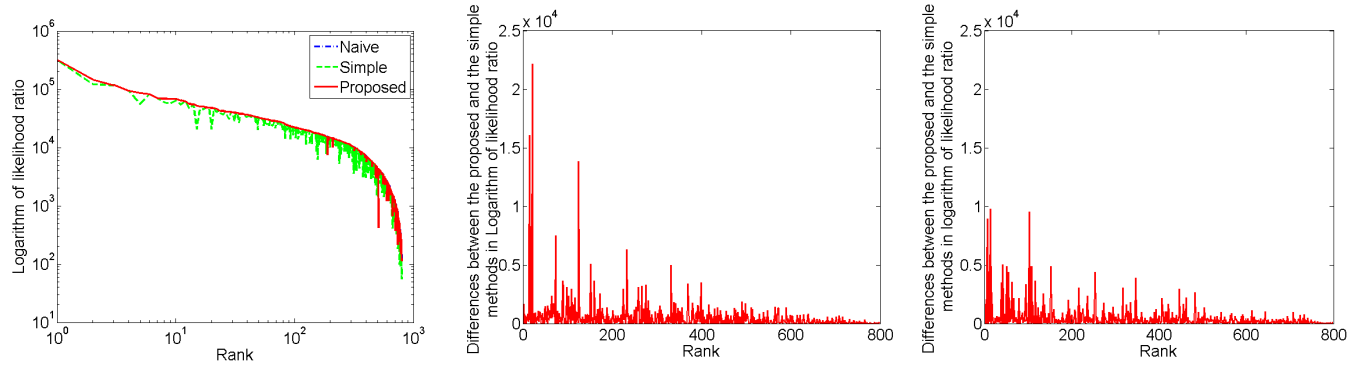
まず，計算効率を比較する為に$J = 2$の場合に各手法が要した計算時間を図 1(a)に示す．図の横軸は拡散系列の系列長であり，縦軸は計算時間（秒）である．この結果から，予想通りナイーブ法が最も多くの計算時間を要しており，対象とする拡散系列長に対してほぼ二乗のオーダーで実際の計算時間も増加していることがわかる．これに対して，シンプル法と提案法にかかる計算時間は非常に短いものとなっており，いずれも拡散系列長に対して線形オーダーで増加していることがわかる．提案法はシンプル法に対して付加的な反復が必要な分だけ計算時間が長くなっているが，その差がそれ



(a) $J = 2$　(b) $J = 3$　(c) $J = 5$

図 1　ナイーブ法，シンプル法，提案法の実行時間の比較

Fig.1　Comparison of computation time among the tree (naïve, simple, and proposed) methods.



(a) $J = 2$　(b) $J = 3$　(c) $J = 5$

図 2　ナイーブ法，シンプル法，提案法の検出精度の比較

Fig.2　Comparison of accuracy among the three (naïve, simple, and proposed) methods.

ほど大きくないことも分かる．実際，提案法による反復回数は平均 2.2 回，最大で 7 回であった．さらに，J が大きくなった場合のシンプル法と提案法の計算時間を比較する為に，$J = 3$ と $J = 5$ の場合の各手法の計算時間を図 1(b)と図 1(c)に示す．これらの場合においても $J = 2$ の場合と同様に，シンプル法，提案法いずれも拡散系列の系列長に対してその計算時間は線形オーダーで増加し，提案法の計算時間はシンプル法の計算時間の高々2.5 倍程度であることがわかる．

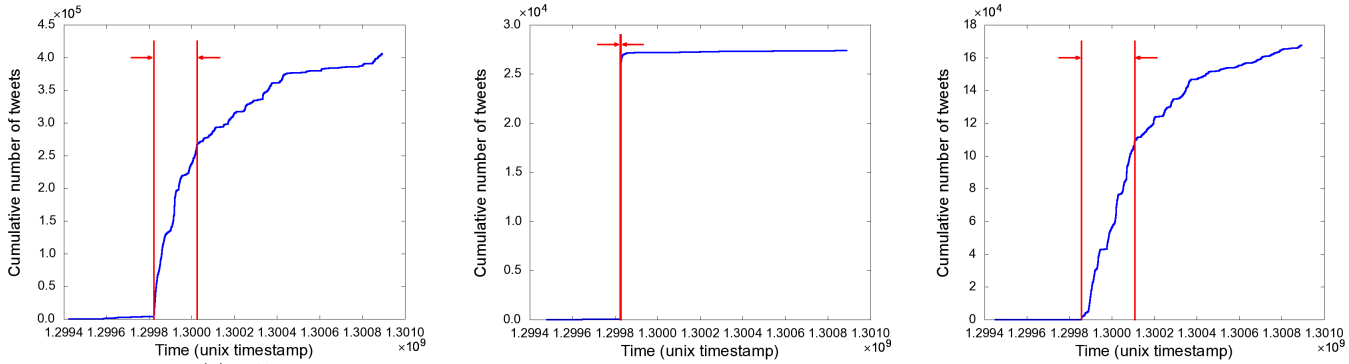
次に，$J = 2$ の場合における変化点検出精度の比較結果を図 2(a)に示す．ここでは，ナイーブ法で得られた解を最適解とし，横軸を各拡散系列をナイーブ法で得た尤度比の降順に並べた場合の順位，縦軸を各拡散系列に対する各手法による解のもつ尤度比の対数としている．この結果は，シンプル法は全体的に尤度比が低く最適解とは異なる変化点を検出することが多いのに対し，ナイーブ法の結果を表す青い線が提案法の結果を表す赤い線にほぼ上書きされて確認できないことからも分かるように，提案法は尤度比が低い場合を除いた多くの場合において最適解を見つけることができていることを示している．提案法の精度が尤度比の低い系列に対して低下するのは，それらの系列におけるバースト期間が明確ではないためであると考えられる．実際には，全 798 系列のうち最適解を得ることができたのはシンプル法の 171 系列（21.4%）に対し，提案法は 713 系列（89.4%）であった．また，ナイーブ法により得た最適解に対する尤度比と各手法で得た解の尤度比の比の平均はシンプル法が 0.881 であったのに対し，提案法は 0.976 となっており，尤度比の観点からも提案法による精度向上が認められる．また，計算時間と同様に，$J = 3, 5$ の場合のシンプル法と提案法の検出精度の比較をそれぞれ図 2(b)，図 2(c)に示す．これらの図では，横軸は対数表示にしていない点を除いて図 2(a)と同様にナイーブ法における尤度比の順位（降順）であるが，縦軸は提案法により得た解の尤度比の対数値からシンプル法により得た解の尤度比の対数値を引いた値を示している．提案法は 4.3 節で述べたようにシンプル法で得られた解より高い尤度比をもつ解を求めるため，この差が大きいほど基準とするナイーブ法の解とシンプル法の解がより大きく乖離していることを意味する．これらの結果から，J の値が大きくなっても上位系列では尤度比の差が大きい傾向は同じであり，全体的に差が小さくなるものの，シンプル法は依然として多くの系列で尤度比がより低い解しか見つけることしかできていないことがわかる．これらの結果から，提案法はシンプル法と比較して僅かに計算時間が増加するものの，変化点（バースト期間）検出精度は大幅に向上することが確認できた．

次に，$J = 2$ の場合に尤度比の対数値が上位 3 位となった拡散系列におけるバースト期間の傾向を調べた．それらの拡散結果に含まれる総ツィート数，およびバースト期間の始点，終点，バースト期間始点付近に見られる主な話題を表 1 にまとめる．また，各拡散結果に対する累積ツィート数の時間変化を図 3 に示す．図の横軸は時間，縦軸は累積ツィート数を表す．グラフ中の縦軸に平行な赤線は提案手法が検出した変化点の位置を表しており，それらに挟まれる区間がバースト期間となる．

まず，表 1 から，地震に関する緊急性の高い内容が短時間のうちに爆発的にリツィートされたことがバースト期間の

表 1　J= 2 の場合の尤度比対数値上位 3 位の情報拡散系列におけるバースト期間先頭 10 件に含まれる主なツィートの内容
Table 1　Major topics appearing among the first 10 tweets of the burst periods of the top 3 diffusion sequences in terms of logarithm of likelihood ratio in case of J= 2.

順位	総ツィート数	バースト期間の始点	バースト期間の終点	バースト期間初期の主な内容
1	450,739	2011/3/11 14:48:13	2011/3/13 23:13:04	NHK 広報局（@NHK_PR）が発信した地震速報のリツィート（10/10）
2	27,372	2011/3/11 15:13:57	2011/3/11 16:19:26	阪神大震災経験者が発信した地震発生時にすべきことのリツィート（10/10）
3	167,528	2011/3/12 00:18:19	2011/3/14 22:08:20	NHK 生活情報部（@nhk_seikatsu）が発信した避難場所での防寒対策のリツィート（10/10）



(a) 1 位　(b) 2 位　(c) 3 位

図 3　J= 2 の場合の尤度比対数値上位 3 位の情報拡散系列における累積ツィート数の変化
Fig.3　Temporal change of cumulative number of tweets in the top 3 diffusion sequences with the highest likelihood ratio in case of J= 2.

始まりとなっていることがわかる．他の系列でも同様の結果を観察しており，このことから，提案手法を用いて検出した変化点に着目する事で，複数の話題を含むツィートの拡散系列データからでも同一の内容の（この場合，緊急性の高い）情報が集中的にリツィートされた区間の始まりを効率的に見つけることができると言える．

一方，表 1 から，2 位の拡散系列の総ツィート数が他と比べて少なく，かつ図 3 からはそのバースト期間が他に比べて非常に短く，前後でツィート数がほとんど変化していないことがわかる．この違いは，2 位の拡散系列が個人のアカウントに対するものであることに起因しているものと考えられる．1 位と 3 位の系列は，それぞれ NHK の広報部と生活情報部のアカウントに対するものであり，それらが発信する情報は日常的にある程度拡散される傾向にある．そのような状況下で表 1 に示すような地震関連の緊急性の高い情報が細かなバースト期間を複数生じさせ，図 3 に見られるような比較的なだらかな累積ツィート数の上昇傾向を示すに至ったと考えられる．実際には，提案手法における変化点数 J の値を大きくすることで，各々のバースト期間を検出することは可能である．これに対して，図 3(b)は，普段はほとんどリツィートされない個人が発信する情報も，地震に関するような緊急性の高いものは極めて短い時間のうちに爆発的に拡散し得ることを示している．

最後に，情報拡散経路を直線状ツリーと仮定した結果と，スター状ツリーと仮定した結果を比較する．1 つのバースト期間のみを含む図 3(b)に対する情報拡散系列に対して，両者を用いて検出した変化点を図 4 に示す．縦軸に平行な赤の実線が前者の結果を，緑の破線が後者の結果を示している．この図から，直線状ツリーを仮定した方が変化点をより精度よく検出できていることがわかる．実際には他の系列でも同様の結果を観察しており，これらから，情報拡散経路として直線状ツリーを仮定することで最尤推定値のよい近似を得ることができると言える．

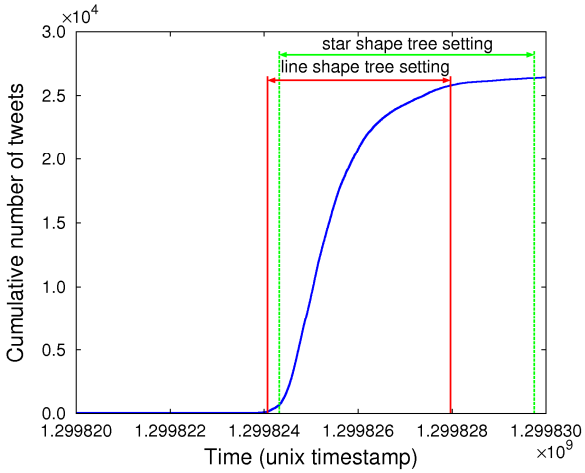


図 4　表 1 の第 2 位の情報拡散系列に対して情報拡散経路として直線状ツリーとスター状ツリーを仮定して検出したバースト期間の比較
Fig.4 Comparison of bursts detected by use of line shape tree setting and star shape tree setting for the 2nd ranked information diffusion sequences in Table 1.

6. おわりに

本論文では，ソーシャルネットワーク上の情報伝搬の遅延時間が指数分布などの確率分布に従うという仮定の下，観測した情報拡散系列からバースト的な拡散現象が生じる区間を検出する問題を検討した．具体的には，指定した時区間ごとの伝搬遅延時間を規定する指数分布のパラメータ系列を観測データに対して最適化することで，それらの区間境界となる変化点を検出する手法を提案した．さらに，Twitter 上を実際に伝搬したツィートデータに提案法を適用し，網羅的探索法に基づくナイーブ法に比べて非常に効率的に，かつ貪欲法に基づくシンプル法に比べて非常に精度よく提案法が変化点を検出できることを実験的に示した．また，実際の情報拡散データの解析を通して，情報拡散系列が複数の話題に関するツィートを含んでいたとしても，検出したバースト期間の開始部分には同一内容のツィートが集中的に含まれる傾向にあること，および実際の情報拡散経路が不明でも直線状ツリーで伝搬経路を近似することで精度よく変化点を検出できることを示した．この結果は，事前に個々の話題ごとにツィート系列を抽出し，その伝搬経路を特定することなく，バースト的に拡散された話題やその期間を提案手法が特定できることを意味しており，その有用性を示すものである．今後の課題としては，ブログ記事等を対象とした既存のバースト検出手法との比較が挙げられる．

【謝辞】

本研究で用いたデータは東京大学 鳥海不二夫氏，NTT 未来ねっと研究所 風間一洋氏によるものであり，その前処理には静岡県立大学の小出明弘氏の協力を得た．また，本研究は科学研究費補助金若手研究(B) (No.23700181)の補助を受けた．

【文献】

[1] Yang, J. and Counts, S.: “Predicting the speed, scale, and range of information diffusion in twitter”, in ICWSM 2010 (2010).
[2] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J.: “Everyone's an influencer: Quantifying influence on twitter”, in WSDM'11, pp.65-74 (2011).
[3] Guille, A. and Hacid, H.: “A predictive model for the temporal dynamics of information diffusion in online social networks”, in WWW'12, pp.1145-1152 (2012).
[4] Katz, L.: “A new status index derived from sociometric analysis”, Sociometry, Vol.18, pp.39-43 (1953).
[5] Bonacichi, P.: “Power and centrality: A family of measures”, Amer. J. Sociol., Vol.92, pp.1170-1182 (1987).
[6] Wasserman, S. and Faust, K.: Social network analysis, Cambridge Univ. Press, Cambridge, UK (1994).
[7] Kempe, D., Kleinberg, J., and Tardos, E.: “Maximizing the spread of influence through a social network”, in KDD2003, pp.137-146 (2003).
[8] Kimura, M., Saito, K., Nakano, R., and Motoda, H.: “Extracting Influential Nodes on a Social Network for Information Diffusion”, Data Min. Knowl. Disc., Vol.20, pp.70-97 (2010).
[9] Goldenberg, J., Libai, B., and Muller, E.: “Talk of the network: A complex systems look at the underlying process of word-of-mouth”, Market. Lett., Vol.12, pp.211-223 (2001).
[10] Watts, D. J.: “A simple model of global cascades on random networks”, PNAS, Vol.99, pp.5766-5771 (2002).
[11] Watts, D. J. and Dodds, P. S.: “Influence, networks, and public opinion formation”, J. Consum. Res., Vol.34, pp.441-458 (2007).
[12] Saito, K., Kimura, M., Ohara, K., and Motoda, H.: “Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis”, in ACML 2009, pp.322-337 (2009).
[13] Saito, K., Kimura, M., Ohara, K., and Motoda, H.: “Selecting information diffusion models over social networks for behavioral analysis”, in ECML PKDD 2010, pp.180-195 (2010).
[14] Sadikov, E., Medina, M., Leskovec, J., and Garcia-Molina, H.: “Correcting for missing data in information cascades”, WSDM2011, pp.55-64 (2011).

大原 剛三 Kouzou OHARA

1995 年大阪大学大学院基礎工学研究科博士前期課程修了．1996 年日本学術振興会特別研究員 DC2. 1997 年大阪大学産業科学研究所助手，同助教を経て，2009 年より青山学院大学理工学部情報テクノロジー学科准教授．データマイニング，機械学習，社会ネットワーク分析の研究に従事. 博士（工学）．IEEE，AAAI，情報処理学会，電子情報通信学会，人工知能学会各会員．

斉藤 和巳 Kazumi SAITO

1985 年慶大・理工・数理科学卒．同年日本電信電話（株）入社．1991 年より 1 年間オタワ大学客員研究員．2007 年より，静岡県立大学経営情報学部教授．機械学習，複雑ネットワーク等の研究に従事．博士（工学）．情報処理学会，電子情報通信学会，日本神経回路学会，日本応用数理学会各会員．

木村 昌弘 Masahiro KIMURA

1987 年阪大・理・数学卒．1989 年同大大学院理学研究科数学専攻修士課程修了．同年，日本電信電話（株）入社．コミュニケーション科学基礎研究所を経て，現在，龍谷大学理工学部電子情報学科教授．複雑ネットワーク科学，データマイニング及び機械学習の研究と教育に従事．博士（理学）．日本数学会，日本応用数理学会，日本神経回路学会，電子情報通信学会各会員．

元田 浩 Hiroshi MOTODA

1965 年東大・工・原子力卒．1967 年同大大学院原子力工学専攻修士課程了. 同年，日立製作所に入社. 同社中央研究所，原子力研究所，エネルギー研究所，基礎研究所を経て 1995 年退社．1996 年大阪大学産業科学研究所教授（知能システム科学研究部門，高次推論研究分野），2006 年定年退職し，現在 現在米国空軍科学技術局アジアオフィス（AFOSR/AOARD）科学顧問．大阪大学名誉教授．大阪大学産業科学研究所招聘教授．原子力システムの設計，運用，診断，制御に関する研究を経て，機械学習，知識獲得，知識発見，データマイニング，社会ネットワーク解析の研究に従事．工博．