情報処理学会論文誌　データベース　Vol.6　No.5　11–22 (Dec. 2013)

情報拡散過程のダイナミクス：非線形モデルの提案と情報予測

松原 靖子[1,a)]　櫻井 保志[1]　B. Aditya Prakash[2]　Lei Li[3]　Christos Faloutsos[4]

受付日 2013年6月21日, 採録日 2013年8月8日

概要：ブログや Twitter をはじめとするソーシャルメディアの発展により，情報交換が活発化し，情報の拡散が非常に速くなっている．本論文では，ソーシャルメディア上における情報拡散と減衰を表現するモデルである SpikeM について述べる．SpikeM は，(a) 情報の拡散過程においてパワー則に基づく減衰パターンを有し，(b) 有限のノード（ユーザ）を仮定し，(c) 周期性を持つ．モデルの利用により，拡散する情報の質やネットワーク規模のような有用な情報を推定することが可能となり，さらに外れ値検出や時系列予測等の実用的なタスクを処理することができる．実データを用いた実験では，提案モデルが効果的に情報拡散のパターンを表現することを示した．

キーワード：情報拡散，ソーシャルネットワーク

Dynamics of Information Diffusion in Social Networks

Yasuko Matsubara[1,a)]　Yasushi Sakurai[1]　B. Aditya Prakash[2]　Lei Li[3]
Christos Faloutsos[4]

Received: June 21, 2013, Accepted: August 8, 2013

Abstract: The recent explosion in the adoption of search engines and new media such as blogs and Twitter have facilitated faster propagation of news and rumors. How quickly does a piece of news spread over these media? Does the rising and falling pattern follow a simple universal law? In this paper, we propose SpikeM, a concise yet flexible analytical model for the rise and fall patterns of influence propagation. Our model has the following advantages: (a) unification power: it generalizes and explains earlier theoretical models and empirical observations; (b) practicality: it matches the observed behavior of diverse sets of real data; (c) parsimony: it requires only a handful of parameters; and (d) usefulness: it enables further analytics tasks such as forecasting, spotting anomalies, and interpretation by reverse-engineering the system parameters of interest (e.g. quality of news, count of interested bloggers, etc.). Using SpikeM, we analyzed 7.2 GB of real data, most of which were collected from the public domain. We have shown that our SpikeM model accurately and succinctly describes all the patterns of the rise-and-fall spikes in these real datasets.

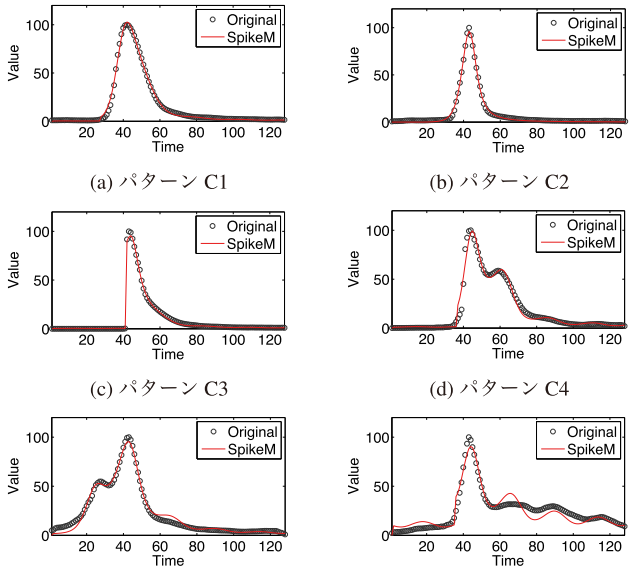
Keywords: information diffusion, social networks

1. まえがき

ブログや Twitter をはじめとするインターネットメディアの普及により，オンライン上でのニュースや噂の伝播速度が増している．あるニュースは緩やかに拡散し，ゆっくりと減衰していく．一方で，別のニュースは急激に伝わると同時に，すぐに消えていく．本論文では，ニュースをはじめとする情報がオンラインメディア上でどのように伝わり，減衰していくかに着目する．

ソーシャルメディア上における情報伝播の解析は非常に重要な問題である．特に，拡散と減衰のダイナミクスの解明は 現在さかんに行われている研究の 1 つである．YouTube

[1] 熊本大学
Kumamoto University, Kumamoto 860–8555, Japan
[2] Virginia Tech., Blacksburg, VA 24060, U.S.A.
[3] University of California, Berkeley, CA 94720–1776, U.S.A.
[4] Carnegie Mellon University, Pittsburgh, PA 15213–3891, U.S.A.
[a)] yasuko@cs.kumamoto-u.ac.jp



(a) パターン C1　(b) パターン C2　(c) パターン C3　(d) パターン C4　(e) パターン C5　(f) パターン C6

図 1　オンラインメディア上に現れる 6 つの代表的な情報拡散パターン（K-SC [43]）と SpikeM の効果

Fig. 1　Modeling power of SpikeM: six types of spikes (K-SC from [43]) in dots, and our model fit in solid red line.

データを用いた先行研究 [6] では，情報拡散には 4 種類（endogenous-subcritical，endogenous-critical，exogenous-subcritical，exogenous-critical）のパターンがあり，ブログデータを用いた文献 [43] では，**図 1** にあげるように，主に 6 種類の拡散パターンが存在すると主張されている．一方，本研究では，まったく別の視点からこの問題に取り組む．我々の手法は，“単一のモデル” によって，これらの様々な情報拡散のパターンを表現する．提案モデルである SpikeM [27] [*1] はシンプル（データセットのサイズと比較し少ない数のパラメータ）であるにもかかわらず，上記のあらゆるダイナミクスを生成することができる．

図 1 は，オンラインメディアにおける 6 つの代表的な情報拡散のパターン（黒点線）と，それに対する SpikeM の学習結果（赤線）である．それぞれのシーケンスは 1 時間ごとの情報の伝播数で，長さは 5 日間である．SpikeM は 7 つのパラメータを使用することによって，6 種類すべての拡散パターンを高精度で表現することができる．

本論文の目的は，ブログ等に代表されるオンライン上の情報拡散過程を，時系列モデルとして表現することである．以下では簡略化のために主にブログの例について言及するが，提案手法は他の様々な拡散過程（購買数の変化，コンピュータウィルスの伝染，Twitter 上での噂の伝播等）においてもモデル化が可能である．本論文で取り組む問題は以下のとおりである．

問題 1　事前情報として，(a) ブログユーザの数 N とネットワーク構造，(b) あるイベントの発生時刻 n_b とそのイベントの質・注目度 β，(c) イベント直後（時刻 n_b）に影響を受けるノード（ユーザ）の数 S_b が与えられたとき，情報拡散過程のダイナミクスを表現する．

もう 1 つの重要な問題は，情報拡散過程のパターンが与えられたうえで，そのダイナミクスを図 1 にあるように，単一のモデルで表現することである．

問題 2　情報拡散過程のパターンが与えられたとき，少ない数のパラメータを用いて，それらのダイナミクスを柔軟に表現する非線形モデルを推定する．

ここで重要な点として，情報拡散の分析において，モデルの各パラメータは直感的な意味を持つことが望ましい．たとえば，ユーザの総数，ニュースの注目度等の意味のある数値をパラメータとして扱いたい．逆にいえば，自己回帰モデル（AR: autoregressive model）のような既存のモデルでは，a_1，a_2 といった，係数値を用いて時系列データを表現するため，パラメータ単体では意味を持たず理想的とはいえない．加えて，情報拡散のモデルはシンプルであるべきである．すなわち，モデルパラメータはできる限り少ない方が望ましい．

表 1　既存手法との比較

Table 1　Capabilities of approaches. Only our approach meets all specs.

	C-S	K-SC	SI	AR	SpikeM
パワー則による減衰	√				√
システム同定			√		√
非線形性			√		√
周期性				√	√
将来予測の能力			√	√	√

1.1　関連研究と本研究の位置づけ

本研究で提案する非線形モデル SpikeM は，(a) 情報の拡散過程においてパワー則に基づく減衰パターンを有し，(b) 有限のノード（ユーザ）を仮定し，(c) 周期性を持つ．従来研究には，これらの特長をすべて持つモデルは存在しない．たとえば，AR や ARIMA に代表される時系列モデルは**線形**であり，**指数則**に基づく減衰パターンを持つため，実データのパターンを表現できない．加えて，無限大に発散する可能性がある．**表 1** は既存研究と SpikeM の能力の比較である．Crane と Sornette らによる C-S 法 [6] は，ノード（ユーザ）の総数に関する有限性を保証していない．K-SC [43] による代表的な拡散パターン（図 1）はパラメトリックなモデルではなく，将来予測の能力を有さない．SI モデルは数理疫学における代表的な非線形モデルであり，マーケティングにおける新製品の拡散過程を表現する Bass モデル [3] に近い性質を持つ．SI モデルはピーク時までの拡散過程とその後の減衰両方において，指数則に従うという性質を持つため，実データにおけるパワー則の減衰パターンを表現できない．

[*1] ソースコード：http://www.cs.cmu.edu/~yasuko/SRC/spikeM.zip

1.2　本論文の貢献

本論文では情報拡散過程の解析ためのモデルとしてSpikeM [27] について述べる．SpikeM は以下のような特長がある．

- 既存研究における情報拡散モデルを一般化し，文献 [20], [43] を含む様々なパターンを表現する．
- 少ないパラメータ数で構成され，それぞれのパラメータが直感的な意味を有する．
- SpikeM の利用により，将来の予測や外れ値検出等の重要なアプリケーションを実現する．

2.　提案モデル

本章では，ネットワーク上における情報拡散過程を表現するモデルである SpikeM について述べる．提案モデルは，図 1 にあげられるような実際の情報拡散過程における特徴を表現する必要がある．その特徴とは，まず (a) 無限大に発散することなく必ず収束することを保証し，(b) 情報拡散はパワー則に基づき減衰していくこと，そして，(c) 周期性を持つことである．(a) については，ネットワーク上のユーザの数を有限とすることで解決できる．(b) については，ネットワークの各ノード（ユーザ）の感染力をべき指数（たとえば –1.5）によって減衰させることで表現することができる．(c) については，ユーザの時間帯ごとの活動の変化を考慮することでモデルをより現実的なものとする．

2.1　SpikeM の概要

ここでは，提案するモデルの概要を示すと同時に，SpikeM を構成するパラメータの説明を行う．提案モデルは，初めに N 人のブログユーザを想定する．この時点では，どのユーザもイベント（ニュース）について知らされておらず，ブログサイトにも記事をポストしていないものとする．続いて，時刻 n_b に特定のイベントが発生する．ここでいうイベントとは，たとえば，インドネシアの津波のニュースの速報や，選挙活動のスピーチに関する話題を指す．すると時刻 n_b においてまず，即座に S_b 人のユーザが自身のブログにポストする．本論文では，このようなイベントを**外部からのショック**と呼ぶ．n_b はイベントの誕生時刻，S_b はイベントの外部からのショックの強さを示す．

さらに，ニュース（イベント）の質，注目度に関するパラメータとして β を用いる．もし $\beta = 0$ だった場合，そのニュースは誰からも注目されていないため，すぐに消えてしまう．β が高い値の場合は，より多くのユーザが興味を持ち，ブログポストを行う．ここで，$\beta * N$ は，初期のイベントの強さ（イベントの影響力）を示す．これは，1 人のユーザがニュースを知らされたときに，次の時刻に何人のユーザが影響されるかを表現する[*2]．

[*2] 厳密には $N-1$ 人であるが，簡略化のために N とする．

SpikeM の中で最も重要な要素は減衰関数 $f(\tau)$ である．これは，時間 τ が経過するにつれて，そのブログポストがどの程度の影響力を与えるかを示す．なお，既存の数理疫学モデルである SI モデルでは，減衰関数 $f()$ を定数であると仮定している．これはたとえば，1 度病気になった患者がそれ以降つねに一定の確率で病気を他者へ感染させるようなケースである．しかし，近年の分析によると，情報拡散における影響力はパワー則に基づき時間とともに減衰していく例が数多く見られている．

提案モデルは以下のような振舞いをする．

- イベント（ニュース）が発生する時刻 n_b までは何も起きない．
- 時刻 n_b にイベントが発生すると同時に，即座に S_b 人のブログユーザがそのニュースに関する記事をブログにポストする．
- 他のブログユーザが，すでにポストされたイベントに関する記事を読み，一定の確率でそのニュースの影響を受ける．その後イベントの影響を受けた（ニュースについて通知された）ユーザは自身のブログ内でそのニュースについて言及（ポスト）する．

本手法ではモデルをシンプルにするため以下の条件も加える．

- ブログユーザは同じニュースに関する記事を 1 度しかポストしない．
- それぞれのイベントは完全独立の関係にあり，1 つのイベントに対し外部ショックは 1 度しか加えられない．

我々の目的は，時刻 n においてブログポストを行ったユーザの数 $\Delta B(n)$ を，時刻 n とその他のパラメータに関する式で表現することである．ここでのパラメータとは，ブログユーザの総数 N やニュースの影響力の強さ β 等である．この時点では，まだ周期性については考慮しない．この基本的なモデルを SpikeM-Base と呼ぶ．

2.2　SpikeM-Base

SpikeM は次にあげられる 2 つの状態のノード（ブログユーザ）から構成される．

- U（Un-informed）：ニュースの内容をまだ通知されていないブログユーザ
- B（Blogged）：ニュースを通知され，その後ブログにニュースの内容をポストしたユーザ

時刻 n の時点でまだそのニュースの影響を受けていないユーザの数を $U(n)$ とする．時刻 n にニュース内容を通知されたユーザの数を $\Delta B(n)$ とする．1 度ニュースについて通知され影響を受けたユーザは，その噂についてただちに（時刻 n に）自分のブログへポストすることとする．

モデル 1（SpikeM-Base）　提案モデル SpikeM-Base は，以下の 2 つの式から構成される．

$$\Delta B(n+1) = U(n) \cdot \sum_{t=n_b}^{n} \big(\Delta B(t) + S(t)\big) \cdot f(n+1-t) + \epsilon \quad (1)$$

$$U(n+1) = U(n) - \Delta B(n+1) \quad (2)$$

ここで，

$$f(\tau) = \beta * \tau^{-1.5} \quad (3)$$

であり，初期状態を次のように設定する．

$$\Delta B(0) = 0, \;\; U(0) = N$$

また，外部ショックとして $S(n)$ を考える．これは，イベントの誕生時刻 n_b に与えられる影響の大きさで，次のような式で表現される．

$$S(n) = \begin{cases} 0 & (n \neq n_b) \\ S_b & (n = n_b) \end{cases} \quad (4)$$

2.2.1　提案モデルの根拠

ここではモデル 1 の細部を考察する．

- 部分式 $\Delta B(t) + S(t)$ は，時刻 t における，ニュースを通知されたユーザの数，外部ショック（ニュースサイトによる報道等）の強さの和である．この部分式が他のユーザに与える影響力（感染力）は，時間の経過とともに減少し，これを減衰関数 $f()$ で表現する．最終的には，イベントの誕生時刻 n_b からのすべての時刻（$t = n_b, \ldots, n$）における部分式の影響力の総和を計算することで，現時刻における全体での影響力を得る．
- 影響力を示す $f()$ は，パワー則に従う．ここでは，先行研究 [2], [21] に基づき，-1.5 のベキ指数を用いる．
- 時刻 n の時点でまだニュースを知らされていない $U(n)$ 人のユーザのうち，新たにニュースの影響を受けるユーザの数は，過去すべての時刻における影響力の総和と $U(n)$ の積により決定される．
- ノイズ ϵ は，イベントとは独立の要因を示す．たとえば，Twitter における #egypt というハッシュタグは，一般的な意味におけるエジプトに関するツイートと，2012 年 11 月にタハリール広場で起きた抗議デモのニュースに関するツイートが独立に存在している．多くの場合において，$\epsilon \simeq 0$ となる．

上記の定義に加えて，$B(n) = \sum_{t=0}^{n} \Delta B(t)$ となり，ブログユーザの総数 N は固定であると考える．つまり，$B(n) + U(n) = N$ である．

2.3　周期性をともなう情報拡散モデル：SpikeM-Full

本章の冒頭で示したとおり，ユーザの活動には周期性をともなうことがある．たとえば，夜中や明け方は，ブログの活動を停止しているユーザが多く，一方で日中には大量のブログポストが発生していると考えられる．このような時間帯ごとの活動の変化は，モデル 1 では表現できない．

モデル 2（SpikeM）　提案モデル（SpikeM）は，以下の式によりブログユーザの周期的な活動を表現する．

$$\Delta B(n+1) = p(n+1) \cdot \Big(U(n) \cdot \sum_{t=n_b}^{n} \big(\Delta B(t) + S(t)\big) \cdot f(n+1-t) + \epsilon\Big) \quad (5)$$

$$p(n) = 1 - \tfrac{1}{2} P_a \Big(sin\big(\tfrac{2\pi}{P_p}(n + P_s)\big) + 1\Big) \quad (6)$$

ここで，$U(n)$，$S(t)$，$f(n)$ は，モデル 1 と同様とする．

モデル 2 はモデル 1 に周期関数 $p(\cdot)$ を加えたものである．このモデルでは，ニュースを受け取る可能性のあるユーザの数 $U(\cdot)$ が時間帯によって変動する．たとえばユーザが就寝している時間帯にはニュースの拡散力が弱まり，少人数しかそのイベントの影響を受けない（感染しない）．周期関数 $p(\cdot)$ は以下で構成される．

- 周期 P_p：24 時間等の周期性を示す．
- 位相 P_s：周期のずれを示す．
- 振幅 P_a：周期性の強弱を示す．

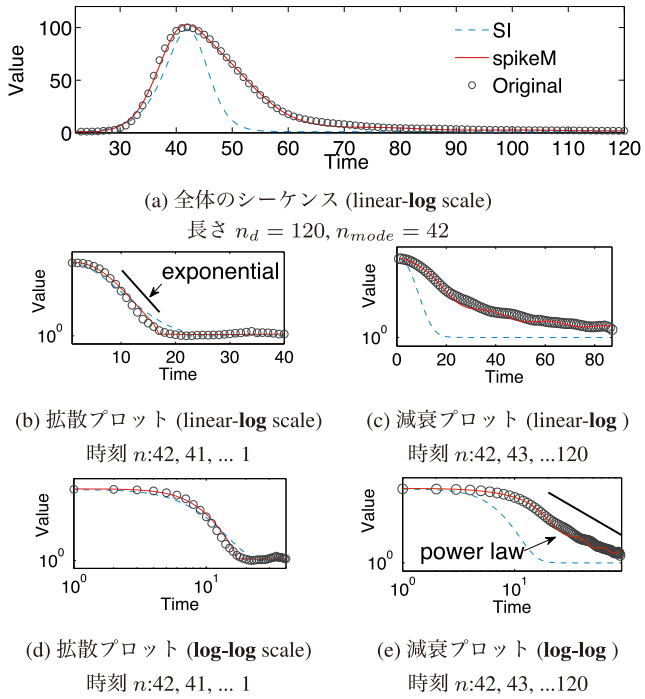
たとえば毎日昼に活動が大きくなる場合，$P_p = 24$ hours，$P_s = 18$ となる．振幅 P_a が大きいと，時間帯によるユーザの活動量に差が出る．周期性が見られない場合，$P_a = 0$ となる．

表 2　主な記号と定義
Table 2　Symbols and definitions.

記号	定義
N	あるイベントに影響されるユーザの潜在的な人数
n_d	シーケンスの長さ
n	時刻（$n = 0, \ldots, n_d$）
$U(n)$	**U**n-informed：ニュースの内容を通知されていないユーザ数
$B(n)$	**B**logged：ニュースの内容を通知されたユーザの総数
$\Delta B(n)$	時刻 n においてニュースを通知されたユーザ数
$f(\tau)$	あるブログポストの τ 時刻後における影響力（感染力）
β	イベント（ニュース）の影響力の強さ
$S(n)$	時刻 n における外部ショックの大きさ
n_b	ニュース速報が報道された時刻（イベントの誕生時刻）
S_b	イベント誕生時刻 n_b における外部ショックの強さ
P_p	周期（1 日周期等）
P_a	周期の振幅
P_s	周期の位相

2.4　モデルパラメータの学習

提案モデルは 7 つのパラメータで構成される：$\boldsymbol{\theta} = \{N, \beta, n_b, S_b, \epsilon, P_a, P_s\}$．実際のシーケンス $X(n)$，$n = 1, \ldots, n_d$ が与えられたとき，本手法は最小二乗法に基づき次式で表現されるエラーの総和が最小になるようなパラメータを選ぶことでモデルの学習を行う：



(a) 全体のシーケンス (linear-**log** scale)
長さ $n_d = 120, n_{mode} = 42$

(b) 拡散プロット (linear-**log** scale)
時刻 n:42, 41, ... 1

(c) 減衰プロット (linear-**log**)
時刻 n:42, 43, ...120

(d) 拡散プロット (**log-log** scale)
時刻 n:42, 41, ... 1

(e) 減衰プロット (**log-log**)
時刻 n:42, 43, ...120

図 2　SpikeM と SI モデルによる図 1（C1）の学習結果の比較．提案手法（赤線）は指数則に基づき拡散し，パワー則で減衰する．一方 SI モデル（青線）は，拡散，減衰ともに指数則に従う

Fig. 2　Fitting results of SpikeM vs. SI for pattern C1 in Fig. 1. The original sequence (in gray circles), and our model (red line) have an exponential rise and a power-law drop.

$$D(X, \boldsymbol{\theta}) = \sum_{n=1}^{n_d} (X(n) - \Delta B(n))^2. \qquad (7)$$

なお，SpikeM は強い非線形性を持つため，モデルの学習にはパラメータの発散に頑健なレーベンバーグ・マルカート法（LM: Levenberg-Marquardt）[22] を用いた．

2.5　モデルの分析—指数則とパワー則

SpikeM は，既存の拡散モデルとは決定的に異なる特長がある．それは，指数則で情報が拡散したのちパワー則で減衰していくことである．本節では，実データを用いて提案モデルの表現能力を分析する．**図 2** は，図 1（パターン C1）の情報拡散過程をそれぞれ SpikeM と SI モデルを用いて学習した結果である．ここで，ブログユーザ $\Delta B()$ の波がピークに達した時刻を n_{mode} とする（この場合 $n_{mode} = 42$）．図 2 (b)，(d) はイベント誕生の時刻 n_b からピーク n_{mode} までの間のシーケンスを x 軸の方向に反転したプロットであり，これを**拡散プロット**と呼ぶ．同様に，図 2 (c)，(e) はピーク n_{mode} から時刻 n までの部分シーケンスを示しており，これを**減衰プロット**と呼ぶ．それぞれ，線形スケール，対数スケールにおいて比較を行った．図に示すとおり，実データは指数則の上昇と，パワー則の減衰パターンを持っており，SpikeM は高い精度でのフィッティングに成功している．一方，既存手法である SI モデルは，上昇，減衰ともに指数則に基づくという性質があるため，拡散プロットにおいては高い精度である一方，減衰時にはうまくパターンを表現していない．

表 3　K-SC（図 1）における 6 つの情報拡散パターンに対する学習モデルのパラメータの値

Table 3　The model parameters of our SpikeM best fitting on six patterns of K-SC (see Fig. 1).

	C1	C2	C3	C4	C5	C6
N	2407	1283	1466	3079	4183	3435
$\beta * N$	0.95	**1.00**	0.86	0.92	0.79	0.69
n_b	26	17	**40**	35	0	34
S_b	4.73	0.06	**114.13**	23.24	2.58	45.58
ϵ	0.36	0.01	0.43	1.48	0.32	**13.97**
P_a	0.18	0.06	0.22	**0.38**	**0.28**	**0.39**
P_s	12	5	7	**6**	**2**	**2**

3.　評価実験

SpikeM の有効性を検証するため，実データを用いた実験を行った．本実験は，以下の諸問題に取り組む．

(1) 実データを用いた提案モデルの精度の検証（Q1-3）

(2) SpikeM を用いた予測精度の検証（Q4）

3.1　データセット

本論文では以下のデータセットに対し評価実験を行った．

- *MemeTracker*[*3]: このデータセットはアメリカ国内のブログにおいて 3 カ月間（2008/8/1–10/31）に発生したフレーズ（meme）を集めたものである．フレーズは，主に政治や時事ニュース等に関連する内容である．本研究ではフレーズの中から特に出現回数の多かった 1,000 個のフレーズとそのフレーズが頻出する 1 週間のシーケンスを選び使用した．
- *Twitter*[*4]: このデータは，Twitter 内で 8 カ月間（2011/6/1–2012/1/1）に現れたハッシュタグを収集したものである．本実験では，700 万以上のツイートの中から，頻出するハッシュタグを 10,000 件選び使用した．
- *GoogleTrends*[*5]: このデータセットは，Google による検索クエリの頻度を集計したものである．各シーケンスが各クエリ（たとえば “tsunami” や “harry potter”）の出現頻度を表す．

3.2　Q1：K-SC クラスタにおけるモデルの精度

このデータセットに関する結果はすでに 1 章の図 1 において示している．図 1 のとおり，SpikeM は，ブログメディアにおける代表的な拡散パターンを高精度で表現している．**表 3** はモデルの学習結果の詳細である．SpikeM

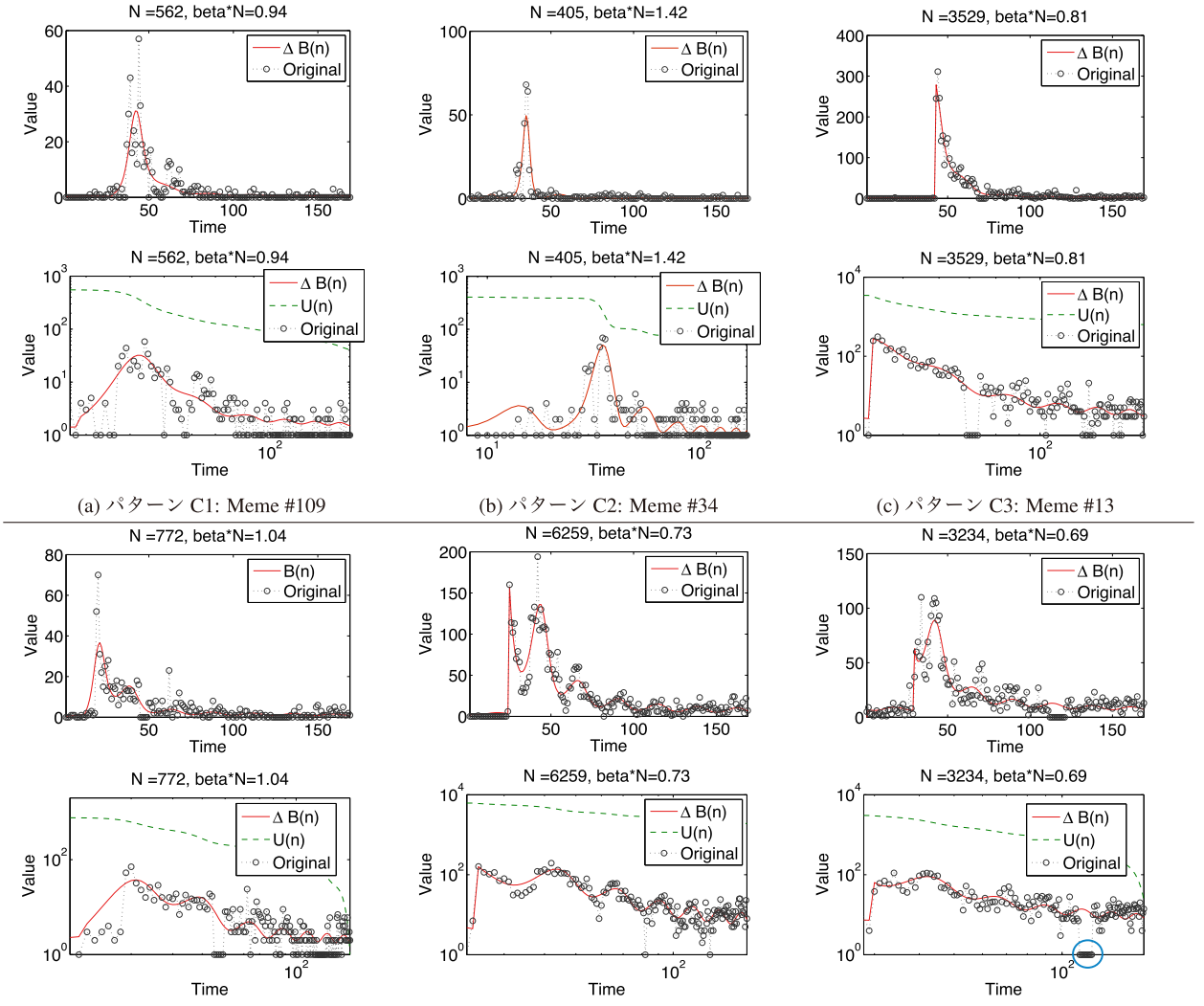
*3　http://memetracker.org/
*4　http://twitter.com/
*5　http://www.google.com/insights/search/

は 7 つのパラメータで表現され，それぞれが拡散パターンの具体的な振舞いを説明している．まず，潜在的なユーザの総数 N はすべてのシーケンスパターンにおいて 2,000〜3,000 の間となっている．これはピーク時の値が 100 になるようにデータを正規化しているためである．続いて，イベントの影響力の値である $\beta * N$ は $0.7 - 1.0$ の間となっている．C1 と C2 は非常に近い性質を持ったパターンであるが，C2 はより強い影響力 $\beta * N = 1.0$ を持っているパターンであることが分かる．外部ショックの強さ S_b については，パターン C3 が時刻 $n_b = 40$ において $S_b = 114$ という高い値を持っている．これは，C3 が他のパターンに比べ，外部からのショックを強く受けていることを意味している．パターン C4，C5，C6 は，強い周期性がある点が特徴的である．これは $P_a \simeq 0.4$ というパラメータによって確認できる．C4 と C5 の違いは位相のずれ（$P_s = 6, 2$）による違いが顕著であり，一方 C6 は強いバックグラウンドノイズ $\epsilon = 13.97$ を含む．このことから，C6 は情報拡散パターンのほかに，独立要因による影響を受けていると考えられる．

表 4 K-SC における SpikeM と SI モデルの精度比較（*RMSE*）
Table 4 Fitting accuracy of SI vs. SpikeM on six patterns of K-SC. SpikeM consistently outperforms SI with respect to accuracy (*RMSE*) between the original values and the models.

Pattern	C1	C2	C3	C4	C5	C6
SpikeM	**1.84**	**1.61**	**0.97**	**4.08**	**3.33**	**5.89**
SI	15.64	6.78	19.65	25.29	20.36	21.76

表 4 は，フィッティングの精度比較である．ここでは二乗平均誤差（*RMSE*）を用いて実データとモデルの比較を行った：$RMSE = \sqrt{\frac{1}{n_d}\sum_n^{n_d}(X(n) - \Delta B(n))^2}$．ここでは

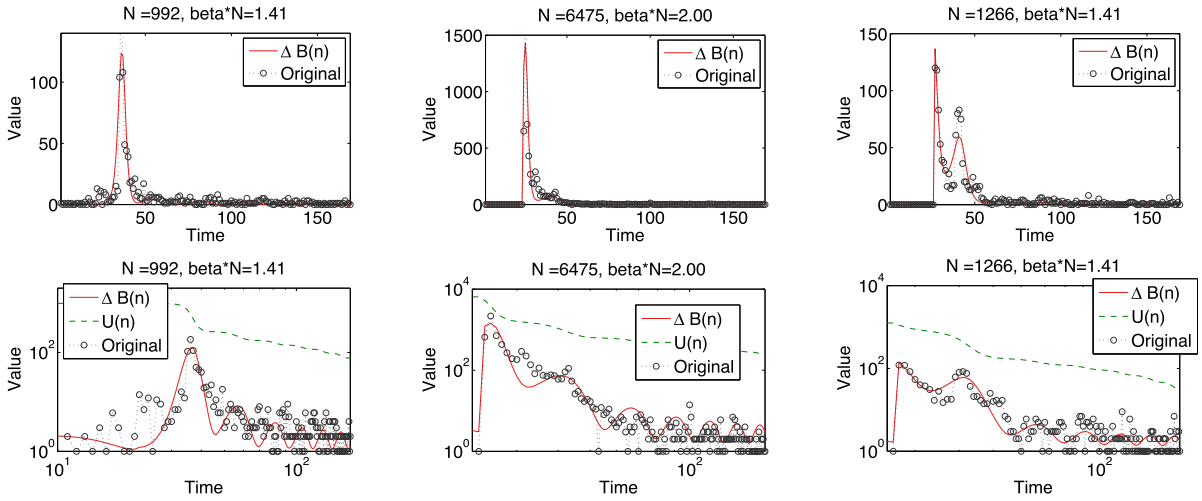


(a) パターン C1: Meme #109　(b) パターン C2: Meme #34　(c) パターン C3: Meme #13

(d) パターン C4: Meme #87　(e) パターン C5: Meme #9　(f) パターン C6: Meme #3

#109	the most serious financial crisis since the great depression	#87	what is required of us now is a new era of responsibility
#34	i love this country too much to let them take over another election	#9	you can put lipstick on a pig
#13	hope over fear, unity of purpose over conflict and discord	#3	yes we can yes we can

図 3 *MemeTracker* データにおける代表的な情報拡散パターンと，SpikeM による学習結果
Fig. 3 Results of SpikeM fitting on six patterns from *MemeTracker* dataset.



(a) #assange　(b) #stevejobs　(c) #arresteddevelopment

図 4　*Twitter* データにおける 3 つのハッシュタグの拡散と，SpikeM の結果
Fig. 4　Results of SpikeM fitting on three hashtags from *Twitter* dataset.

SpikeM と既存手法である SI モデルを比較した．2 章の図 2 で比較したとおり，SI モデルはパワー則に基づく減衰パターンを有しないため，正しくパターンを表現できない．一方，我々の提案モデルは高い精度で盛り上がりと減衰の両パターンを正しく表現できることが分かる．

3.3　Q2：*MemeTracker* データとモデルの比較

図 3 は *MemeTracker* データにおける提案モデルの学習結果（赤線）である．ここでは，K-SC で発見された 6 つの代表的なクラスタ（C1–C6）に類似するシーケンスを選び，学習結果を示した．表は各シーケンスに関連するフレーズ（meme）であり，これらのフレーズはすべて 2008 年におけるアメリカの政治ニュースに関連する内容である．各シーケンスはそれぞれ線形スケール（上段），対数スケール（下段）で示している．提案手法である SpikeM は，パワー則に基づく拡散パターンを柔軟に表現することができる．特に，図中の対数スケールにおいて，提案モデルは実データの減衰パターンを正しくとらえていることが確認できる．以下では，各パターンについて詳細を考察する．

- パターン C1，C2：潜在的なユーザの総数 $N \simeq 500$ はほぼ同様であるが，パターン C2 は C1 よりも急激に情報が拡散している（情報拡散の強さは $\beta * N = 1.4$）．
- パターン C3：強い外部ショックと緩やかな減衰を持ち，微かに周期性を帯びている．
- パターン C4，C5：明確な周期性を持つ．パターン C5 の “lipstick on a pig” は 6 つのシーケンスの中で潜在的ユーザの総数が最も多い（$N = 6259$）．
- パターン C6：フレーズ “yes we can” は，全体的に一定数の周期的なノイズも観測できるが，それと同時に時刻 $n = 40$ 付近に強い情報拡散パターンが発生している．これは，ブログユーザが日常的なフレーズとして “yes we can” を利用しているのと同時に，Barack Obama の選挙活動におけるスローガンの意味としてこのフレーズをブログ内でポストしているためと考えられる．加えて，時刻 $n = 120$ 付近には外れ値が観測されるが，我々の手法は情報拡散の本質をとらえることができるためノイズに頑健であり，局所的なノイズの影響を受けない．

3.4　Q3：その他のデータにおけるモデルの検証

3.4.1　*Twitter* データ上の情報拡散

図 4 は，*Twitter* データのハッシュタグにおけるモデルの学習結果である．*Twitter* は *MemeTracker* データに非常によく似た拡散パターンを持つことが分かる．ここでは頻出する 3 つのハッシュタグについてのみ結果を示す：(a) #assange：このハッシュタグは WikiLeaks の創立者 Julian Assange に関するニュースの拡散過程である．まず初めに少人数による言及があり，その後ピーク時（2011/12/5）に向かって情報が伝搬しているのが分かる．(b) #stevejobs：Steve Jobs の死去（2011/10/5）直後にニュースが発生し，急激にユーザに伝わると同時に，長い期間にわたって話題が広がっている．(c) #arresteddevelopment：これは映画 “Arrested Development” に関する話題である．明確な周期性がある．

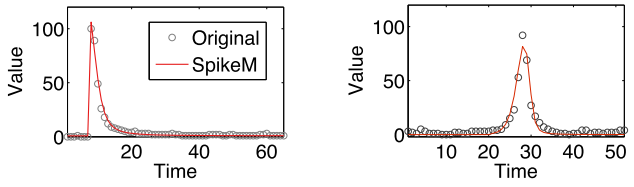
3.4.2　*GoogleTrends* データ上の情報拡散

情報拡散のパターンは，Blog や *Twitter* をはじめとするソーシャルネットワークだけではなく，Google 等の検索エンジン内のクエリにおいても観測される．**図 5** は，*GoogleTrends* データにおける情報拡散パターンの例である．図 5 (a) は外部からの強いショックが要因となり発生したイベントの例（2005 年に発生したインドネシアの大地震に関するニュース）である．一方，図 5 (b) は 2007 年に公開されたハリーポッターの映画が，口コミ等によって広

まっていく過程が示されている．図に示すとおり，2 つの異なる性質を持つパターンを SpikeM は正しく表現している．

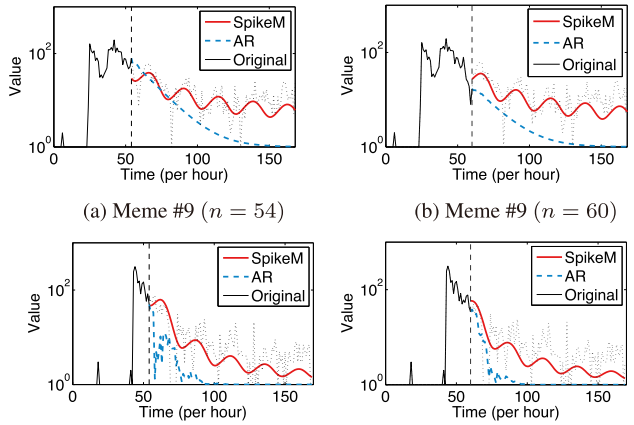
3.5　Q4：SpikeM を用いた拡散過程の予測

上に示したとおり，SpikeM は様々な拡散パターンを表現することができる．これを利用することで，将来の拡散過程を予測することができる．**図 6** は，*MemeTracker* データ（図 3 Meme #9，#13）を用いた予測結果である．まず初めに，時刻 $n = 54$ および $n = 60$ までのシーケンス（黒線）を用いてそれぞれモデルの学習を行う．その後学習したモデルを用いて以降 5 日間のパターン（赤線）を予測した．提案モデルの精度を比較するため，既存手法である AR（青線）の予測結果も示した．AR の係数は SpikeM のパラメータ数と同様の 7 に設定した．図に示すとおり，AR が長期的な予測に失敗しているのに対し，提案手法は正しく将来の拡散過程を予測している．より具体的には，2 つのシーケンスに対する SpikeM の予測時のエラーがそれぞれ $RMSE = 9.26$，8.93 だったのに対し，AR は $RMSE = 13.98$，14.19 となった．ここでさらに重要なことに，SpikeM は，上記のようなピーク時以降のパターンだけではなく，イベントの発生直後からピーク時にかけての拡散過程も予測することができる．これに関しては 4.1 節で詳しく述べる．



(a) “tsunami” (2005)　(b) “Harry Potter” (2007)

図 5　*GoogleTrends* データにおけるモデルの学習精度（赤線）
Fig. 5　SpikeM fitting on *GoogleTrends* dataset: the volume of searches for the keyword (in black dots) and fitting results (in red lines). Note that the window size is per week.



(a) Meme #9 ($n = 54$)　(b) Meme #9 ($n = 60$)

(c) Meme #13 ($n = 54$)　(d) Meme #13 ($n = 60$)

図 6　*MemeTracker* データにおける情報拡散過程の予測
Fig. 6　Results of tail-part forecasting on *MemeTracker* data.

4.　アプリケーション

これまで述べたように，提案手法である SpikeM では直感的な意味を持つパラメータを推定することができる．本章では，提案手法およびそれらのパラメータの活用方法について議論すると同時に，実用的なアプリケーション例として代表的な 3 つについて紹介する．

4.1　“What-if” シナリオと将来予測

3.5 節においてすでに SpikeM の予測能力について紹介したが，ここではさらに重要なアプリケーションとして，“what-if” シナリオにおける将来予測問題に取り組む．具体的には，3.5 節の予測問題がシーケンスのピーク時までのパターンを用いて，その後の動きを予測していたのに対し，この “what-if” シナリオではピーク以前のダイナミクスも予測する．これは容易に解決できる問題ではない．なぜなら，一般的な情報拡散過程では，非常に短い期間の間にピーク点に達してしまうからである．しかしこれが繰返しのあるイベントであればどうだろうか．この “what-if” シナリオでは，単体のイベントに対して予測をするのではなく，過去に起きたイベントの情報を用いることで，次のイベントの拡散過程を予測する．ここで，ハリーポッターの映画を例として考える．もしハリーポッターの映画の続編が近日公開されるとしたら，ソーシャルメディア上でどのように話題が盛り上がるかを事前に予測できるだろうか．**図 7**（*GoogleTrends*）を用いて問題を定義する．まず，事前情報として，(a) 2009 年に発表された過去の映画 “Harry Potter and the Half-Blood Prince”（$n = 185$）の拡散パターンと (b) 次に発表予定となっている続編 2 つの公開日時（水色矢印で示される点，時刻 $n = 255$，289），(c) 公開日時以前 8〜2 週間前のアクセス数（水色線）が与えられるとする．ここでの目標は，これらの情報から続編 2 つの情報拡散過程を推定し，ピーク点を予測することである．

4.1.1　解決法

SpikeM は，N の値からハリーポッターに興味を持つ

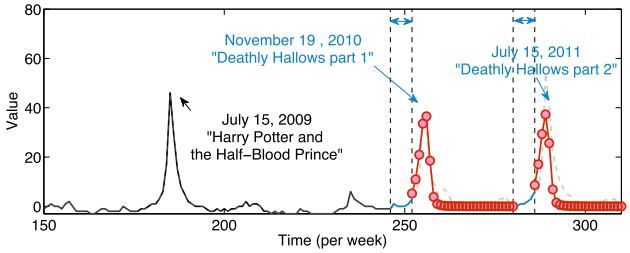


図 7　映画ハリーポッター（*GoogleTrends*）における情報拡散パターンの予測
Fig. 7　Results of “what-if” forecasting for the Harry Potter series.

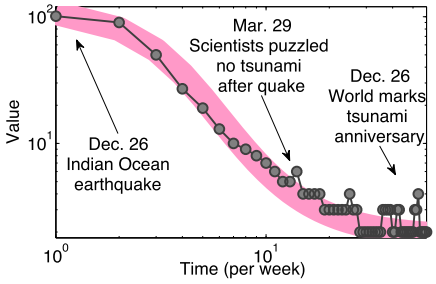


図 8　*GoogleTrends* データにおける外れ値検出
Fig. 8　Outlier detection on *GoogleTrends* dataset.

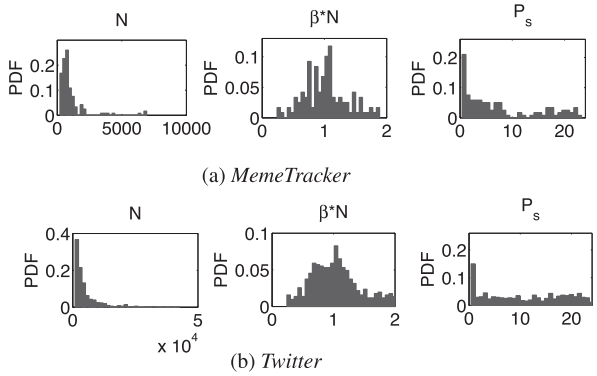


図 9　*MemeTracker* と *Twitter* データにおけるパラメータ（N, $\beta * N$, P_s）の分布の様子
Fig. 9　Reverse engineering: pdf of three parameters: $N, \beta * N, P_s$ over 1,000 memes/hashtags. (a) *MemeTracker*: total potential bloggers $N \simeq 1,000$, and strength of "first burst" $\beta * N \simeq 1.0$. More than 90% of the memes have clear daily periodicity with high activities around 6 pm (i.e., $P_s \simeq 0$). (b) *Twitter*: similar trends except more spread in P_s, possibly, due to multiple time zone.

可能性のあるユーザの数を，β の値からこのイベントの影響力（口コミによる情報伝搬の強さ）を推定することができる．そこで本手法では，すべてのイベント（ハリーポッターシリーズ）の情報拡散過程においてこれらの潜在的なパラメータを共通化することを考え，それぞれのイベント間の違いは，外部ショックの強さ (n_b, S_b) のみであるとする．以下で処理の流れを示す．

(1) まず初めのピーク点前後のパターン（黒線）を用いてパラメータ $\boldsymbol{\theta}$ を学習する．
(2) 学習した $\boldsymbol{\theta}$ を固定したうえで，水色線（$n = 250, 280$）で示す部分シーケンスを用い，2 つの新たなイベントに対する外部ショックの強さ (n_b, S_b) をイベントそれぞれに対し推定する．
(3) 固定したパラメータ $\boldsymbol{\theta}$ と新たに推定した (n_b, S_b) を用い，シーケンスの予測を行う（赤線）．

以上のような処理を用いることで，図 7 に示すとおり，本手法は 2 つの新たなパターンを予測するとともに，ピークの位置も正しく推定することに成功した．

4.2　外れ値検出

3 章で示したように，SpikeM は実データのパターンを高い精度で表現することができる．これを利用することで，外れ値検出問題を容易に実現することができる．より具体的には，時刻 n におけるデータと推定モデルの距離を対数スケールを用いて次式で計算する：$\delta(n) = log|X(n) - \Delta B(n)|$．この値 $\delta(n)$ が閾値を超える場合にその時刻 n を外れ値とする．**図 8** は，図 5 (a) に示した津波のシーケンスを対数スケールでプロットしたものである．ここで，黒点はオリジナルの値，赤線は SpikeM のフィッティング結果を表す．図中において，いくつかの外れ値が確認できる．たとえば，(a) 3 月 29 日にこのイベントとは異なる別の地震が発生している．さらに，(b) 12 月 26 日（$n = 52$）に非常に大きなスパイクがあるが，これはこの大地震が発生したちょうど 1 年後の地震の記念日にあたる．

4.3　リバースエンジニアリング

提案モデルの最大の強みは，構成する各パラメータが，イベントに興味を持つユーザの潜在的な人数やニュースの注目度等の直感的な意味を持つことである．この特長を利用することで，観測される拡散パターンから，本来のデータの振舞いを理解することができる．**図 9** は，*MemeTracker* と *Twitter* の 2 つのデータセットから，SpikeM を用いてそれぞれ 1,000 以上の拡散パターンを学習し，各パラメータ（N, $\beta * N$, P_s）の分布を示したものである．以下では得られた知見について議論する．

知見 1　あるイベントに興味を持つ潜在的なユーザ数 N は，両データにおいて主に $N = 1,000 - 2,000$ である．

知見 2　イベント生成直後の影響力の強さは両データにおいて，およそ $\beta * N \simeq 1.0$ に分布している．

上記の 2 つの知見は，*MemeTracker* と *Twitter* の両データにおいて情報拡散過程に共通の振舞いがあることを意味する．つまり，Blog と Twitter が類似したソーシャルアクティビティを持つことを示唆している．

知見 3　両データにおいて，24 時間単位の周期的な活動が見られる．特に *MemeTracker* データの (a) 周期の位相は $P_s = 0$ であり，早朝の活動が小さく，夕方に活動が増加する一方で，*Twitter* データでは (b) 位相 P_s に幅がある．

実験では，両データセットにおいて 90%以上のシーケンスに，1 日単位の周期性が見つかった．周期性に関する特徴として，*Twitter* データには様々な位相 P_s を持つシーケンスが観測された．これは，*MemeTracker* データがアメリカ国内のユーザであるのに対し，*Twitter* のユーザが様々なタイムゾーン（アメリカ，イギリス，オーストラリア，インド等）を使用しているためである．

5.　関連研究

関連研究は以下の 3 つに分類される．

5.1　時系列データ解析

時系列データの解析に関する研究は様々な分野で進められている [4]．自己回帰モデル（AR: autoregressive model），線形動的システム（LDS: linear dynamical systems），カルマンフィルタ（KF: Kalman filters）は代表的な技術であり，これらに基づく時系列の解析と予測手法が数多く提案されている [12], [23], [24], [25]．しかしこれらの技術はすべて線形という特徴を持つ．非線形のモデルにおいては最近傍探索（NNS: nearest neighbor search）[5] や人工ニューラルネットワーク（ANN: artificial neural networks）[41] を用いる手法が主流であり，予測モデルとしての表現能力は不十分である．時系列シーケンスを対象とした類似検索やパターン発見問題も様々な研究が行われているが [7], [8], [13], [15], [26], [28], [30], [31], [33], [36], [37], [40]，時系列データのバースト性の発見とそのモデル化に着目していない．

5.2　影響伝播

ソーシャルネットワーク上の情報伝播は，SI モデル（susceptible-infected model）に代表される感染症疫学モデル [1] と深い関係性がある．文献 [29] において，情報の減衰パターンがパワー則の性質を持つことが報告されている．より具体的には，ブログにおける単位時間あたりの情報の影響力は −1.5 のベキ指数に基づき減衰する性質を持つ．Barabasi は通信における応答時間の分布がベキ指数 −1 から −1.5 の間のパワー則に従うという性質を発見している [2]．ブログ，ソーシャルメディアの解析および情報の伝播とカスケードに関する研究は様々なものがあり [9], [10], [11], [14], [17], [19], [20], [34], [35], [39], [42]，さらにその発展研究として，情報拡散の発起点を探す研究 [18], [38] も進められている．

5.3　バースト検知

時系列データを対象としたバースト検知は Kleinberg [16]，Zhu ら [44]，Parikh ら [32] に代表される様々な手法が存在するが，これらはすべてネットワーク上のバーストを表現するモデルではない．

6.　むすび

本論文では，オンラインメディア上の情報拡散過程を表現するモデルとして，SpikeM について述べた．SpikeM は既存の知見（K-SC，SI モデル）を一般化すると同時に，パワー則に基づく減衰パターンを含む，様々な実データ上の情報拡散の振舞いを表現する．SpikeM は少ない数のパラメータで構成され，これらのパラメータの利用することにより，‘what-if’ シナリオによる予測や，リバースエンジニアリング等の有用なアプリケーションを実現した．今後の課題として，外部ショックが複数箇所から発生する状況や，より複雑なネットワーク構造を考慮したモデルについて検討していく予定である．

参考文献

[1] Anderson, R.M. and May, R.M.: *Infectious Diseases of Humans*, Oxford University Press (1991).
[2] Barabasi, A.L.: The Origin of Bursts and Heavy Tails in Human Dynamics, *Nature*, Vol.435 (2005).
[3] Bass, F.M.: A New Product Growth for Model Consumer Durables, *Management Science*, Vol.15, No.5, pp.215–227 (1969).
[4] Box, G.E., Jenkins, G.M. and Reinsel, G.C.: *Time Series Analysis: Forecasting and Control, 3rd edition*, Prentice Hall, Englewood Cliffs, NJ (1994).
[5] Chakrabarti, D. and Faloutsos, C.: F4: Large-scale Automated Forecasting Using Fractals, *CIKM* (2002).
[6] Crane, R. and Sornette, D.: Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System, *PNAS* (2008).
[7] Faloutsos, C., Ranganathan, M. and Manolopoulos, Y.: Fast Subsequence Matching in Time-series Databases, *SIGMOD*, pp.419–429 (1994).
[8] Gilbert, A.C., Kotidis, Y., Muthukrishnan, S. and Strauss, M.: Surfing Wavelets on Streams: One-pass Summaries for Approximate Aggregate Queries, *VLDB*, pp.79–88 (2001).
[9] Goetz, M., Leskovec, J., McGlohon, M. and Faloutsos, C.: Modeling Blog Dynamics, *ICWSM* (2009).
[10] Gruhl, D., Liben-Nowell, D., Guha, R. and Tomkins, A.: Information Diffusion through Blogspace, *SIGKDD Explor. Newsl.*, Vol.6, No.2, pp.43–52 (2004).
[11] Guha, R., Kumar, R., Raghavan, P. and Tomkins, A.: Propagation of Trust and Distrust, *WWW*, pp.403–412 (2004).
[12] Jain, A., Chang, E.Y. and Wang, Y.-F.: Adaptive Stream Resource Management Using Kalman Filters, *SIGMOD*, pp.11–22 (2004).
[13] Kahveci, T. and Singh, A.K.: An Efficient Index Structure for String Databases, *Proc. VLDB*, pp.351–360 (Sep. 2001).
[14] Kempe, D., Kleinberg, J. and Tardos, E.: Maximizing the Spread of Influence through a Social Network, *KDD* (2003).
[15] Keogh, E.J., Palpanas, T., Zordan, V.B., Gunopulos, D. and Cardle, M.: Indexing Large Human-motion Databases, *VLDB*, pp.780–791 (2004).
[16] Kleinberg, J.M.: Bursty and Hierarchical Structure in Streams, *KDD*, pp.91–101 (2002).
[17] Kumar, R., Mahdian, M. and McGlohon, M.: Dynamics of Conversations, *SIGKDD*, pp.553–562 (2010).
[18] Lappas, T., Terzi, E., Gunopulos, D. and Mannila, H.: Finding Effectors in Social Networks, *KDD*, pp.1059–1068 (2010).
[19] Leskovec, J., Adamic, L.A. and Huberman, B.A.: The Dynamics of Viral Marketing, *TWEB*, Vol.1, No.1 (2007).
[20] Leskovec, J., Backstrom, L. and Kleinberg, J.M.: Meme-tracking and the Dynamics of the News Cycle, *KDD*, pp.497–506 (2009).
[21] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N.S. and Hurst, M.: Patterns of Cascading Behavior in Large Blog Graphs, *SDM* (2007).
[22] Levenberg, K.: A Method for the Solution of Certain

Non-linear Problems in Least Squares, *Quarterly Journal of Applied Mathmatics*, Vol.II, No.2, pp.164–168 (1944).
[23] Li, L., Liang, C.-J.M., Liu, J., Nath, S., Terzis, A. and Faloutsos, C.: Thermocast: A Cyber-physical Forecasting Model for Data Centers, *KDD* (2011).
[24] Li, L., McCann, J., Pollard, N. and Faloutsos, C.: Dynammo: Mining and Summarization of Coevolving Sequences with Missing Values, *KDD* (2009).
[25] Li, L. and Prakash, B.A.: Time Series Clustering: Complex is Simpler!, *ICML* (2011).
[26] Lin, J., Keogh, E.J., Lonardi, S., Lankford, J.P. and Nystrom, D.M.: Visually Mining and Monitoring Massive Time Series, *KDD*, pp.460–469 (2004).
[27] Matsubara, Y., Sakurai, Y., Prakash, B.A., Li, L. and Faloutsos, C.: Rise and Fall Patterns of Information Diffusion: Model and Implications, *KDD*, pp.6–14 (2012).
[28] Matsubara, Y., Sakurai, Y. and Yoshikawa, M.: Scalable Algorithms for Distribution Search, *ICDM*, pp.347–356 (2009).
[29] McGlohon, M., Leskovec, J., Faloutsos, C., Hurst, M. and Glance, N.: Finding Patterns in Blog Shapes and Blog Evolution, *International Conference on Weblogs and Social Media*, Boulder, Colo. (March 2007).
[30] Papadimitriou, S. and Yu, P.S.: Optimal Multi-scale Patterns in Time Series Streams, *SIGMOD Conference*, pp.647–658 (2006).
[31] Papapetrou, P., Athitsos, V., Potamias, M., Kollios, G. and Gunopulos, D.: Embedding-based Subsequence Matching in Time-series Databases, *ACM Trans. Database Syst.*, Vol.36, No.3, p.17 (2011).
[32] Parikh, N. and Sundaresan, N.: Scalable and Near Real-time Burst Detection from Ecommerce Queries, *KDD*, pp.972–980 (2008).
[33] Patel, P., Keogh, E.J., Lin, J. and Lonardi, S.: Mining Motifs in Massive Time Series Databases, *Proc. ICDM*, pp.370–377 (2002).
[34] Prakash, B.A., Beutel, A., Rosenfeld, R. and Faloutsos, C.: Winner Takes All: Competing Viruses or Ideas on Fair-play Networks, *WWW*, pp.1037–1046 (2012).
[35] Prakash, B.A., Chakrabarti, D., Faloutsos, M., Valler, N. and Faloutsos, C.: Threshold Conditions for Arbitrary Cascade Models on Arbitrary Networks, *ICDM* (2011).
[36] Sakurai, Y., Faloutsos, C. and Yamamuro, M.: Stream Monitoring under the Time Warping Distance, *Proc. 23rd International Conference on Data Engineering, ICDE 2007*, April 15-20, 2007, The Marmara Hotel, Istanbul, Turkey, pp.1046–1055 (2007).
[37] Sakurai, Y., Papadimitriou, S. and Faloutsos, C.: BRAID: Stream Mining through Group Lag Correlations, *SIGMOD Conference*, Baltimore, MD, USA, pp.599–610 (2005).
[38] Shah, D. and Zaman, T.: Rumors in a Network: Who's the Culprit?, *IEEE Trans. Information Theory*, Vol.57, No.8, pp.5163–5181 (2011).
[39] Tong, H., Prakash, B.A., Tsourakakis, C.E., Eliassi-Rad, T., Faloutsos, C. and Chau, D.H.: On the Vulnerability of Large Graphs, *ICDM* (2010).
[40] Vlachos, M., Kozat, S.S. and Yu, P.S.: Optimal Distance Bounds on Time-series Data, *SDM*, pp.109–120 (2009).
[41] Weigend, A.S. and Gerschenfeld, N.A.: *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison Wesley (1994).
[42] Yang, J. and Leskovec, J.: Modeling Information Diffusion in Implicit Networks, *ICDM*, pp.599–608 (2010).
[43] Yang, J. and Leskovec, J.: Patterns of Temporal Variation in Online Media, *WSDM*, pp.177–186 (2011).
[44] Zhu, Y. and Shasha, D.: Efficient Elastic Burst Detection in Data Streams, *KDD*, pp.336–345 (2003).

松原 靖子

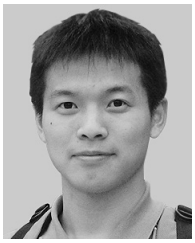
2006 年お茶の水女子大学理学部情報科学科卒業．2009 年同大学大学院人間文化創成科学研究科理学専攻博士前期課程修了．2012 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（工学）．2011～2012 年カーネギーメロン大学客員研究員．データストリーム処理，大規模データマイニングに関する研究に従事．日本データベース学会会員．

櫻井 保志　（正会員）

1991 年同志社大学工学部電気工学科卒業．1991 年日本電信電話（株）入社．1999 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．2004～2005 年カーネギーメロン大学客員研究員．2013 年熊本大学大学院自然科学研究科教授．本会平成 18 年度長尾真記念特別賞，本会平成 16 年度および平成 19 年度論文賞，電子情報通信学会平成 19 年度論文賞，日本データベース学会上林奨励賞，ACM KDD best paper awards（2008，2010）等受賞．データマイニング，データストリーム処理，センサデータ処理，Web 情報解析技術の研究に従事．ACM，電子情報通信学会，日本データベース学会各会員．

B. Aditya Prakash

B. Aditya Prakash is an Assistant Professor in the Computer Science Department at Virginia Tech. He graduated with a Ph.D. from the Computer Science Department at Carnegie Mellon University in 2012 and got his B.Tech. (in CS) from the Indian Institute of Technology (IIT)—Bombay in 2007. He has published more than 25 refereed papers in major venues, holds two U.S. patents, and has given two tutorials (VLDB 2012 and ECML/PKDD 2012). His work has received one best paper award and two best-of-conference selections (CIKM 2012, ICDM 2012, ICDM 2011). His interests include Data Mining, Applied Machine Learning, and Databases, with emphasis on large real-world networks and time series.



Lei Li

Lei Li is a Post-Doctoral researcher at EECS department of UC Berkeley and visiting researcher at CMU. His research interest lies in the intersection of machine learning, statistical inference and database systems. Specifically, he has been working on Bayesian inference in open universe probabilistic models, probabilistic programming language, large-scale learning, time series, communication and social networks. He received his B.S. in Computer Science and Engineering from Shanghai Jiao Tong University in 2006 and Ph.D. in Computer Science from Carnegie Mellon University in 2011, respectively. His dissertation work on fast algorithms for mining coevolving time series was awarded ACM KDD best dissertation (runner up).

Christos Faloutsos

Christos Faloutsos is a Professor at Carnegie Mellon University. He has received the Presidential Young Investigator Award by the National Science Foundation (1989), the Research Contributions Award in ICDM 2006, the SIGKDD Innovations Award (2010), seventeen best paper awards (including two 'test of time'), and four teaching awards. He has served as a member of the executive committee of SIGKDD; he is an ACM Fellow; he has published over 200 refereed articles, 11 book chapters, and one monograph. He holds five patents and he has given over 30 tutorials and over 10 invited distinguished lectures. His research interests include data mining for graphs and streams, fractals, database performance, and indexing for multimedia and bioinformatics data.

（担当編集委員　比戸 将平）