情報処理学会論文誌 Vol.58 No.6 1287–1299 (June 2017)

バースト現象におけるトピック分析

鳥海 不二夫[1,a)] 榊 剛史[1,2]

受付日 2016年10月3日, 採録日 2017年3月3日

概要：近年，ソーシャルメディアにおいては，震災や選挙，炎上などの社会的イベントにより，特定の話題が大きく取り上げられるバースト現象が頻繁に発生している．そのような社会的イベントがどのように社会に受け止められているかを正確に理解するためには，バースト現象を分析し，どのような人々がどのような意見を表明しているかを明らかにするための技術が必要不可欠である．本研究ではバースト現象発生時に，(1) どのようなトピックが含まれるか，(2) 各トピックがどのようなユーザによって拡散されているかを分析することで，バースト現象の詳細を明らかにする手法を提案した．まず，予備評価実験で，提案するトピック分類手法およびユーザ分類手法により適切な結果が得られることを示した．事例分析では提案手法を用いて炎上や自然災害など 5 つのバースト事例の分析を行い，それぞれの事例において，投稿数が多くかつ多様なユーザに語られていたトピック，投稿数が多いが一部のユーザのみに語られていたトピックを明らかにした．本提案手法は，必ずしも新しい手法ではなく，基本的には既存の手法の組合せによるものである．しかしながら，それによってバースト現象の詳細を分析することが可能であることを示した点が本論文における最も大きな貢献である．

キーワード：バースト現象，トピック分類，ソーシャルメディア，Twitter

Topic Analysis for Burst Phenomena

Fujio Toriumi[1,a)] Takeshi Sakaki[1,2]

Received: October 3, 2016, Accepted: March 3, 2017

Abstract: Recently, burst phenomena, which mean specific topics are referred widely, occur frequently on social media. Those are caused by some social events such as natural disasters, public election and flaming. To understand how such social events have been received by the society as a whole, we need a new method, which reveal who has what opinion by analyzing burst phenomenon. In this paper, we propose a method to reveal details of bust phenomenon by analyzing (1) what kinds of topics are included in a burst (2) who propagate each topics. In preliminary evaluation experiments, we show that it is possible to acquire appropriate results by the proposed method for topic clustering and use community detection. In detailed case analysis, we analyze 5 bust cases, including natural disasters and flaming, and uncover that some big topics are referred by various users and other big topics are referred by partial users in each cases. The proposed method basically consists of common techniques; those are not so new one technically. The contribution of this paper is to prove that it is possible to analyze burst phenomena in detail by the combination of existing simple methods.

Keywords: burst, topic classification, social media, Twitter

1. はじめに

ソーシャルメディアにおいては，特定の話題が大きく取り上げられるバースト現象 [1] が頻繁に発生する．バースト現象は，震災や選挙，あるいは炎上といった社会的イベントによって引き起こされる．たとえば，先の東日本大震

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[2] 株式会社ホットリンク
Hotto Link Inc., Chiyoda, Tokyo 102–0071, Japan
[a)] tori@sys.t.u-tokyo.ac.jp

災発生時にはソーシャルメディアの 1 つである Twitter 上では多くの投稿が行われ，その数は通常時の数倍～数十倍にのぼっていた [2] ことが知られている．

バースト現象が発生している状態では，多数の人々が対象事象についてソーシャルメディア上に投稿・拡散するため，多種多様な視点からの情報があふれかえることになる．たとえば，震災時であれば，被災者による投稿，救援者による投稿，それ以外のユーザによる投稿など多岐にわたる．選挙前のバーストであれば，支持する政党や政策によって異なる視点からの投稿が存在する．また，炎上の場合は炎上対象を攻撃するような投稿と，擁護するような投稿が入り交じることも多い．

このとき，どのような人々がどのような意見を表明しているかを理解することは，社会的イベントがどのように社会に受け止められているかを正確に理解するうえで必要不可欠である．たとえば，ある特定の商品の不具合や問題点がソーシャルメディア上で大きな話題となったとしよう．このような企業にとって不利益な情報がバーストすることを炎上と呼ぶが，炎上しているのが商品のユーザなのか，単に炎上を楽しんでいる人たちなのかを理解することは，企業のリスクマネジメントにおいて非常に重要である．どのような立場の人たちがどのような意見を表明しているのかを俯瞰的に理解し，バーストの全体像の把握を行わなければ，誤った対処をしてしまう危険性が高まる．

そこで，本研究ではバースト現象発生時に，

(1) どのようなトピックが含まれるか
(2) 各トピックがどのようなユーザによって拡散されているか

を分析することで，バースト現象の詳細を明らかにする手法を提案する．

バースト現象発生時には，ネット上の様々なメディアで情報が拡散されるためそのすべてを把握し分析することが望ましい．しかしながら，すべてのメディア上でどのようなユーザにどの程度見られたかを把握することは難しい．そのため，各ユーザが明示的に興味を持ったコンテンツを示した情報を用いる必要がある．ユーザが興味を持った対象を取得できるメディアとしては，ソーシャルブックマークやソーシャルメディアなどがある．本研究では，「ユーザが興味を持った内容はソーシャルメディアを通じて拡散しようとする」という仮説に基づき，ソーシャルメディアの 1 つである Twitter[*1] を用いる．

本研究は，計算社会科学の視点から，バースト現象が社会に与える影響をデータから正しく理解できるようにすることを目的としている．これによって，災害事例であればどのような情報をどのような人に提供すればよいかが明らかになる．また，炎上事例であれば対処の方法を明らかにでき，社会問題であればどのような視点で社会がとらえているかを明らかにできるようになるだろう．なお，計算社会科学は学際的な分野であり多様な立場が含まれるが，本論文では，「従来は社会科学で扱われてきた現象を，情報工学の手法を用いて定量的に観測・分析しようとする立場」という意味で用いることとする．

[*1] http://twitter.com

2. 関連研究

ツイートを分類する研究には，García-Silva ら [3] のベクトル空間法と LDA を用いる手法がある．Zhao ら [4] はユーザごとにトピックを持つとする Twitter-LDA モデルを使用して，Twitter からトピックを抽出した．Ramage ら [5] は Labeled LDA を用いてツイートの特徴付けを行っている．また，O'Connor ら [6] の，TweetMotif を用いたツイートの分類がある．Tumasjan ら [7] は，政党や政治家に言及したツイートを抽出することで，Twitter が政治に対する議論フォーラムとして機能していることを示した．

これらの Twitter からの情報抽出手法は，テキストマイニングによるものであるため，URL を含むツイートや短いツイートを扱うことが難しいという問題がある．ツイートそのものの言語情報を用いない手法としては，Rosa ら [8] は，ハッシュタグを用いたツイートの自動分類や，Davidov ら [9] は，ハッシュタグと顔文字を感情の教師データとしたツイートの感情分類がある．また，Baba ら [10] はリツイート行動に基づくツイートの分類を行っている．

ソーシャルメディアからユーザコミュニティを抽出する試みは数多く行われている [11]．その中で，本研究と同様にユーザの分類を目的とした研究としては，下記の研究があげられる．金川らはリンク構造を用いた重複クラスタリング手法によるユーザのクラスタリングを行い [12]，Mano らはユーザごとの記事嗜好を用いたユーザのクラスタリング [13] を行った．

このように，ツイートとユーザを分類する手法はいくつか存在するが，それらを組み合わせることで，バースト状態のトピックについて詳細に分析しようとした研究はない．

3. ツイートとユーザの分類

本章では，バースト発生時に関連するトピックと拡散させたユーザを抽出し，バースト現象に対する社会全体の態度を明らかにする手法を提案する．提案手法の概要は以下のとおりである．

まず，バーストに関連する情報を，拡散したユーザに基づいて分類を行う．具体的には，各投稿に対しリツイート[*2] したユーザの類似性をもとにツイートネットワークを構築した後，そのネットワークをクラスタリングすることで各情報を話題ごとに分類する．

[*2] 自分が受け取った投稿を自分の投稿を見ているユーザに共有する機能

次に，投稿・拡散したユーザをコミュニティごとに分類する．具体的には，「ソーシャルメディア上で頻繁にコミュニケーションを行うユーザ同士は同一コミュニティに所属する可能性が高い」という仮定のもと，各ユーザの相互メンション関係からネットワークを構築しコミュニティ抽出を行う．各コミュニティに含まれるユーザの属性は，コミュニティメンバのユーザプロフィールに特徴的に出現する語を用いて表現する．

3.1 ツイートの分類

Twitterにおける情報の拡散はリツイートによって行われる．まず，これらのリツイートされたツイートについて，含まれる情報ごとに分類を行う．分類には自然言語を用いた分類手法 [14] もあるが，Twitterの投稿は140文字以内という制限があるため，単純な分類は困難であることが知られている．

そこで，本研究では2つのツイートをリツイートしたユーザの重複率に基づいて2つのツイートの類似性を評価し，類似ツイート間にリンクを張ることによってネットワークを構築し，得られたネットワークにクラスタリング手法 [10] を適用することで，情報の分類を行う．

ある2つのツイートを同時にリツイートしたユーザが複数人いた場合，2つのツイートは共通した内容を有していると考えられる．そこで，リツイートしたユーザの重複度から類似したツイートを見つけることが可能である．このように，ツイート関係のみを利用してクラスタリングを行うことで，言語的なクラスタリングでは得られない「興味を示したユーザの類似性」によってツイートを分類する．

ツイート間類似度の算出 ネットワークを構築するにあたり，2つのツイート間の類似度を定義する．前述のとおり，2つのツイートをリツイートしたユーザの重複率を類似度として用いる．

まず，ツイートとユーザ間の関係性を行列 R を構築する．

$$R = [r_{l,m}]$$

$$r_{l,m} = \begin{cases} 1 & (\text{ユーザ } m \text{ がツイート } l \text{ を RT した}) \\ 0 & \begin{pmatrix}\text{ユーザ } m \text{ がツイート } l \text{ を} \\ \text{RT していない}\end{pmatrix} \end{cases}$$

R において，ツイート i に対応する行ベクトルをツイートの特徴ベクトル $\boldsymbol{t}_i$ とする．なお，U はデータセット内における全ユーザ集合を表す．

$$\boldsymbol{t}_i = (r_{i,0}, r_{i,1}, \ldots, r_{i,|U|})$$

このとき2つのツイート t_i，t_j の類似度をSimpson係数を用いて式 (1) のように定義する

$$Sim(t_i, t_j) = \frac{|\boldsymbol{t}_i \cdot \boldsymbol{t}_j|}{min(|\boldsymbol{t}_i|, |\boldsymbol{t}_j|)} \quad (1)$$

なお，このような類似度を計る指標としては，Simpson係数のほかにJaccard係数，Dice係数などがあるが，共起を用いた関係性の強さを表現するための指標としてはSimpson係数が適切であるとされている [15]．

ネットワークの構築 データセットに含まれるツイート集合からネットワークを構築する．ここでは，前述の類似度を手がかりに，類似度の上位 N_{th} 件にあたるツイートペアの間にリンクを張り，重みあり無向ネットワークを構築する．

ある人物が2つのツイートを拡散した場合，その2つのツイートには何らかの共通性があるといえよう．ただし，1人だけが同時にリツイートしただけでは，その共通性はきわめて個人的なものかもしれない．しかし，もしそれが複数の人物によってなされたのであれば，一般性のある共通性があると考えられる．そのため，類似度が高い，つまりユーザの重複率が十分高ければ，2つのツイートには何らかの意味で共通性があると判断できる．

このようにして，共通性のあるツイートどうしにリンクを張ることによってネットワークを構築すれば，共通性に基づくネットワークを構築することができる．ネットワークを構築する手順は下記のとおり．

(1) ツイート集合 T から，2つのツイート t_i，t_j を取り出し，類似度 $Sim(t_i, t_j)$ を算出する．これをすべてのツイートペアについて行う．
(2) 得られた全ペアのうち，類似度が上位 N_{th} 件に含まれるツイートペアを抽出する．
(3) 抽出したツイートペアの間にリンクを張る．

ネットワーククラスタリング 得られたネットワークについて，コミュニティ抽出を行い，関係性の深いツイートの集合を獲得する．類似度 $Sim(t_i, t_j)$ を重みとしてツイートの類似性を示すネットワークを構築し，得られたネットワークに対してコミュニティ抽出に用いられるネットワーククラスタリングの手法を適用する．コミュニティ抽出には，モジュラリティ [16] Q を基準とするLouvain法 [17] を用いた．モジュラリティとは各クラスタの結合度合いを表す指標であり，この値が最大になるようにクラスタリングを行うことで，より結合度の高いクラスタが得られる．

3.2 ユーザコミュニティ抽出手法

バースト現象の分析において，「誰がそのトピックに興味を持っているか」を明らかにするために，ユーザの分類を行う．

従来の社会学やマーケティングなどにおいては，現象の俯瞰を目的としたユーザ分類として，地域，年代，職業といったデモグラフィクスが用いられてきた．同じデモグラフィクスを持つ個人は同じような興味・価値観を持つとい

う仮定のもと，このようなアプローチが用いられてきた．確かに Web の登場以前は，情報流通の手段が限られており，同じデモグラフィクスを持つ個人間で，日常的に情報が流通することが多かったたため，「同じデモグラフィクスを持つ個人の興味・価値観は類似している」という仮定は成立していたと考えられる．たとえば，大学生であれば，地元の年齢が近い知り合いと日常的にコミュニケーションをしていることは容易に想像される．

しかし，Web およびソーシャルメディアの登場以降，そのような地域間コミュニケーションや年代間コミュニケーションに対する障壁が小さくなり，そういったデモグラフィクスの壁を超えて，自身の趣味や興味が類似した個人とコミュニケーションを行う機会が増加したことが推測される．そこで，本研究では，ソーシャルメディア上のコミュニケーションデータから，日常的にコミュニケーションしているコミュニティを抽出し，それをバーストトピックを分析するためのユーザの分類として用いる．

具体的には，対象となるトピックについてツイートを行ったユーザを，ユーザ同士のコミュニケーション行動に基づいてコミュニティに分割する．コミュニティへの分割は，ユーザ全体をネットワークととらえたうえで，コミュニティ抽出に用いられるネットワーククラスタリングの手法を適用する．また各コミュニティを特徴づけるために，各コミュニティに特徴的な語を抽出する．

コミュニティの抽出

まずユーザ集合において，ユーザのコミュニケーション関係を表すネットワークを構築する．具体的にはある一定期間に相互にメンションしあっている 2 ユーザをコミュニケーション関係にあるユーザペアと見なし，そのようなユーザ間にリンクを作成することでコミュニケーションに基づくユーザネットワークを構築する．その後，コミュニケーションネットワークに Louvain 法を適用することで，コミュニティ抽出を行う．

ユーザネットワークの構築 ユーザの相互メンション関係からネットワークを構築する．メンションとはあるユーザから特定のユーザ宛に送られる投稿のことである．ここでは，ユーザ A とユーザ B がお互いにお互い宛てのメンションを N_{mt} 以上投稿している場合に，ユーザ A，ユーザ B 間にリンクを張る．これを全ユーザペアについて行うことで，重みなし無向ネットワークを構築する．

ユーザネットワークの構築において，対象データと同じ期間のデータから相互メンション関係を抽出するのが適切であるように思える．しかし，バースト事例によりユーザ間の関係性にも変化が生じる可能性がある．実際，東日本大震災の場合には大規模にユーザの Twitter 利用方法が変わったことが報告されている [18]．そこで，本論文では，相互メンション関係の抽出は，分析対象データの直前の期間のデータを用いることとする．具体的には，分析対象となるデータの前々月～前月の 2 カ月間の日本語 10%ランダムサンプリングデータから，相互メンション関係を抽出する．本研究では，「あるイベントが社会に受け止められているかを正確に理解する」ということを目的としているため，ユーザネットワークにはより多くのユーザが含まれることが望ましい．ここで 2016 年 5 月～6 月における相互メンション頻度の度数分布表を**図 1** に示す．図 1 より，相互メンションの頻度が 1 回の場合が全体の約 50%を占めているため，$N_{mt} \geq 2$ とすると，多くのユーザがユーザネットワークに含まれなくなってしまう．そこで相互メンションの閾値を $N_{mt} = 1$ とした．

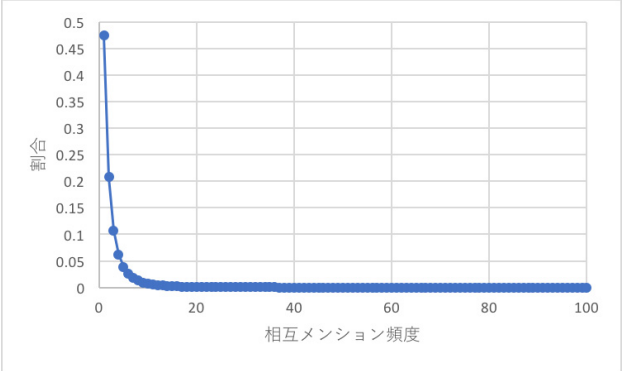

図 1 2016 年 5 月～6 月における相互メンション頻度の度数分布表
Fig. 1 Frequency distribution of inter-mentions frequency in May to June, 2016.

コミュニティの分割 次に構築したネットワークをコミュニティに分割し，それをユーザ分類として用いる．コミュニティ分割手法には，トピックのクラスタリングと同様に Louvain 法を利用する．ただし，情報の分類とは異なり，重みなしのネットワークであるため，重みなしの Louvain 法を適用する．ユーザの分類においては重みなしネットワークを用いる理由は下記である．まず，ユーザ分類において，重みあり相互メンションネットワークにクラスタリングを適用することでユーザコミュニティを抽出することを考える．このような場合，ユーザ間のコミュニケーション頻度が高いほど親密であることが知られているため [19]，ユーザ間の親密度が高い，いわば「強い」コミュニティが抽出されると推測される．一方，本研究においては，「あるイベントが社会に受け止められているかを正確に理解する」という観点から興味の近さや居住地の近さといったユーザの社会的属性に利用可能な，一定の共通性を持った「弱い」ユーザコミュニティを抽出することを目的としている．このような目的においては，親密度のような強い関係性ではなく，「コミュニケーションがあるかないか」のような弱い関係性を用いる

ことが望ましい．そこで，本研究では，重みなしの相互メンションネットワークに対して，重みを考慮しない Louvain 法を適用し，上記のような一定の共通性を持ったユーザコミュニティを抽出することを目指す．なお，上記のような弱いコミュニティを抽出する目的においては，本来的には相互フォロワ関係を用いることが望ましいが，Twitter のフォロー関係にはタイムスタンプがない（フォロー関係がいつ発生したか分からない），大規模にデータを取得することが困難であるといった観点から，本研究では用いないものとする．

コミュニティ特徴語の抽出

抽出したコミュニティにどのような人が含まれているかを理解するために，コミュニティを特徴づけるような特徴語を抽出する．各ユーザを特徴づけるテキストとして，Twitter においてはユーザの投稿とユーザプロフィールに含まれる自己紹介文が候補として考えられるが，本研究では，自己紹介文を用いる．これは，下記のような理由による．Twitter の投稿には文字数制限はあるものの投稿数には制限がなく，多様な投稿が可能であるため，挨拶や引用記事タイトルなど，ユーザの特徴とは無関係な文が多数含まれてしまう．一方，自己紹介文は最大 160 文字という文字数制限があるため，ユーザの特徴抽出という観点から見た場合にノイズが含まれている可能性が低いと考えられる．

各ユーザの自己紹介文を収集した後，各コミュニティを構成するユーザの自己紹介文を結合し，それを 1 文書とする．このように生成したコミュニティを特徴づける文書集合において，各コミュニティの文書ごとに語の tf-idf（term frequency - inversed document frequency）値を算出する．そして各コミュニティの文書ごとに，tf-idf 値の上位 50 語を，コミュニティ特徴語とした．ここでは特徴語として名詞のみを用いるものとする．また，$tf < 3$ であるような低頻度語，および $df > 0.2N_{all}$（N_{all}：全コミュニティ数）となるような高頻度語は特徴語としては用いないこととする．

さらに特徴語群からコミュニティについてラベルを付与する．具体的には，Wikipedia の全記事データをインポートした全文検索エンジンを用意する．各コミュニティについて，特徴語の上位 50 語を検索クエリとして全文検索エンジン上で検索を行い，関連度が高い文書の記事タイトルをコミュニティのラベルとして用いる．全文検索エンジンは Elasticsearch Version1.5，分かち書きエンジンは Kuromoji，分かち書きエンジンの辞書は MeCab-neologd，文書の関連度算出アルゴリズムは tf-idf を用いた．

4. 災害情報バースト事例におけるトピック分析

4.1 データセット

本章では実際のバースト事例として，2014 年に発生した御嶽山の噴火に関するツイートデータを用い，災害情報バースト時のトピック分析を行った．

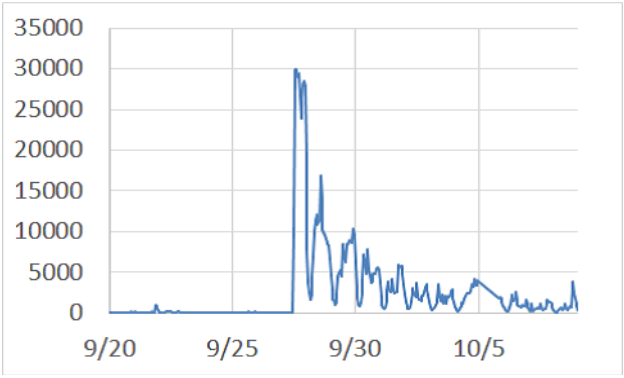

図 2 御嶽山噴火データにおけるツイート数の時系列変化
Fig. 2 Numbers of Tweets in Ontake eruption data.

この災害事例では，一部の被災者が噴火の直前や直後にツイートを行っていたため，それらのツイートが話題となるとともに，戦後最大の火山災害ということで多くの情報が飛び交った．ここでは，どのようなトピックが存在し，どのような人々によってそれらのトピックが拡散されたのかを分析し，御嶽山の噴火がソーシャルメディアでどのように扱われたのかを明らかにする．

本事例の概要は付録 A.1.1 に示す．ここでは，309,638 ユーザによる 1,097,091 ツイートを収集し，データとして用いた．**図 2** は，データ収集期間に得られたツイート数の 1 時間ごとの変化を示したものである．

4.2 トピック分類

4.2.1 パラメータ設定

まず，3 章で提案した手法を用いて，本バースト事例に含まれるツイートを分類した．

ここでは，ある程度のユーザに拡散した情報のみを扱うものとして，最低リツイート数 k を設定した．また，関係性が少ない 2 つのツイート，すなわち Simpson 係数が小さいリンクの影響を排除するため，Simpson 係数上位 N_{th} までのリンクのみを利用することとした．すなわち，最低リツイート数 k と採用リンク数 N_{th} が分類のパラメータとなる．ここでは，$k = \{10, 20, 50, 100\}$，$100 \leq N_{th} \leq 100{,}000$ で変化させた．

このとき，クラスタリング結果を評価するために，モジュラリティ Q を用いた．これは，不適切なパラメータを用いた場合，トピックが適切に分割できなくなるであろうためである．逆に，トピックが適切に分割できるのであれば，得られたトピックには意味があると考えられる．

クラスタリングの結果の評価を**図 3** に示す．これより，$k = 20$，$N_{th} = 1{,}330$ において $Q = 0.945$ と最大値となった．

このとき，合計 385,527 ツイートから 355 個のトピックが抽出された．

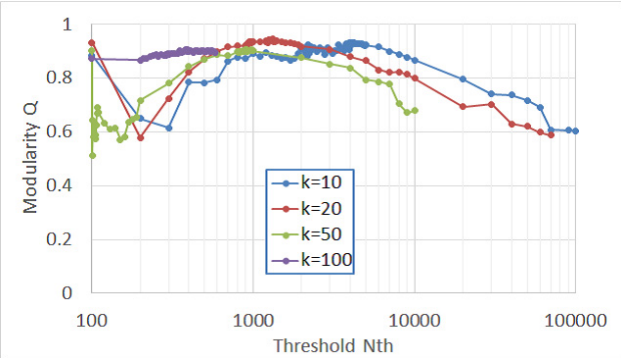

図 3 n, k によるモジュラリティ Q の変化（御嶽山噴火データ）
Fig. 3 Changes in modularity Q for n, k (Ontake eruption).

表 1 抽出されたトピック（トップ 10）
Table 1 Top 10 topics.

	Topics	Teewts	Retweets
T1	被災者の安否	7	28,205
T2	災害救助と科学技術	5	16,335
T3	被災者の報告	12	12,847
T4	噴火と歴史	2	12,262
T5	ニュースとネタ	3	11,874
T6	写真と火山情報	19	9,326
T7	被災者の報告	3	7,587
T8	防災・災害予知	32	7,329
T9	自衛隊の活動	4	7,077
T10	オカルト	7	7,011

得られたトピックのうち，リツイート数の多い方から 10 個のトピックについて**表 1** に示す．

4.2.2 評価実験

ここで，得られたトピックの正しさを評価するために，評価実験を行った．

トピック内にあるすべてのツイートの類似性に関して絶対的な評価を下すことは困難であるため，一対比較による評価を行った．本実験において被験者は「あるツイートとより類似しているツイートはどちらか」を判断する．

まず，任意の 2 つのトピック T_i，T_j をランダムに選択し，T_i から 2 つのツイート t_{i1}，t_{i2} を，T_j からツイート t_j を抽出する．このとき，ツイート t_{i1} を提示ツイートとし，t_{i2}，t_j を選択対象ツイートとした．被験者は，ツイート t_{i1} を読み，t_{i2}，t_j のうちツイート t_{i1} とより類似していると判断した方のツイートを選択する．このような選択を 1 被験者あたり 100 回行った．被験者は 7 名で同一の 100 問の選択を行った．**図 4** は実験画面の例である．上部に配置したツイート（t_{i1}）と類似していると思われるツイートを下部の 2 つのツイート（t_{i2}，t_j）から選択する．なお，t_{i2}，t_j の配置順はランダムに決定された．

また，比較のため言語を用いたクラスタリング手法である LDA [20] によって抽出されたトピックについても，同様の実験を行った．この際，条件を同じにするため LDA

図 4 WEB による評価実験画面
Fig. 4 Evaluation experiment.

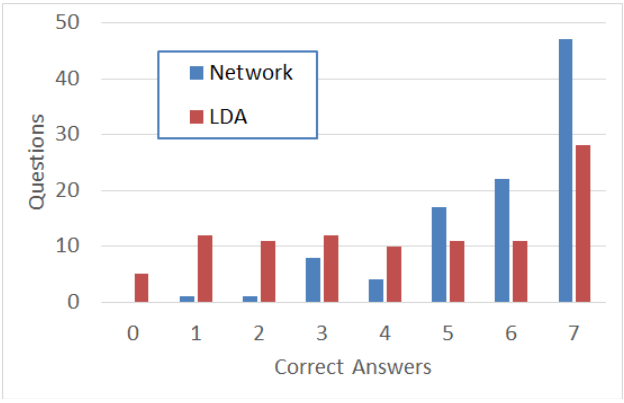

図 5 ツイート類似性判定実験の正解数
Fig. 5 Results of evaluation experiment.

でのトピック数，および総ツイート数は，提案手法で得られたトピックのものと同一になるようにした．なお，LDA に用いた単語は名詞のみとした．

被験者実験の結果を**図 5** に示す．横軸が問題正解者数，縦軸が問題数である．まず，提案手法では全員一致で正しい選択を行った（正解数 7）が 48%であった．一方で，過半数以上（正解数 4 以上）が正解だった場合を正しく分類できていると判断すると，正解率は 90%となり，提案手法によって分類されたツイートは，高い精度で正しく分類できているといえる．

一方，LDA によるクラスタリングでは，全員一致で正しい選択を行った（正解数 7）が 28%であった．（正解数 4 以上）が正解だった場合を正しく分類できていると判断すると，正解率は 60%である．

これより，提案手法は言語を用いたクラスタリングよりも高い精度でトピック抽出が可能であることが示された．

4.3 コミュニティ抽出

4.3.1 抽出結果

トピックの分類と同様に提案手法を用いて各バースト事例に関わったユーザ群からコミュニティを抽出した．

前述のとおり，バースト事例が発生した前々月～前月，つまり 2015 年 7 月 1 日～8 月 31 日の日本語 10%サンプリングデータをすべて用いて，ユーザの相互メンションネッ

表 2 得られたコミュニティとラベル，特徴語（トップ 10）
Table 2 Community, label, and key words.

ID	ユーザ数	ラベル（Wikipedia タイトル）		上位特徴語群				
1	2,658	声優ユニット	水樹奈々	アニメ	奈々	水樹	声優	ライブ
2	2,607	バイク川崎バイク	ばくおん!!	バイク	乗り	自転車	愛車	ロード
3	2,255	チェルノブイリ原子力発電所事故	原子力発電	原発	福島	放射能	tpp	被曝
4	2,134	応援団	オリンピックサッカー日本代表選手	サッカー	応援	fc	サポ	観戦
5	1,959	日本会議	日本のヘイトスピーチ	日本	反日	安倍	保守	支持
6	1,637	ゲーム	ボーダーブレイク	lov	ゲーム	ゲー	ガン	スト
7	1,626	Pixiv	創作版画	創作	成人	注意	同人	ゲーム
8	1,542	Pixiv	ピクシブ（企業）	東方	pixiv	member	net	サークル
9	1,392	Pixiv	ポケットモンスター（アニメ）	ケモノ	気軽	ポケモン	アイコ	ケモナー
10	1,325	アニメ	アニメ+	アニメ	好き	ボカロ	気軽	ゲーム

トワークを構築した後，それに重みなし Louvain 法を用いて，コミュニティへの分割を行う．その後，コミュニティに含まれるユーザのプロフィール文から，各コミュニティを表す特徴語群を取得した後，ラベルを付与する．各事例の分析においては，実際にバーストに関わったユーザと各ユーザが所属するコミュニティを紐付けて，バースト事例におけるコミュニティの分布を作成する．

御嶽山噴火の事例においては，2,322 個のコミュニティが抽出された．コミュニティのユーザ数が多い順に上位 10 件を**表 2** に示す．これより御嶽山の事例には，声優ファンやバイクファン，政治に興味が強いコミュニティ，サッカーファンやゲーム，イラスト描き趣味など，様々なコミュニティのユーザが関わっていることが分かる．

4.3.2 評価実験

提案手法により，得られたコミュニティが適切なものになっているかを評価を行う．

表 2 を見ると，確かに各トピックの特徴語には一定の意味的なまとまりがあるように見える．コミュニティの抽出に用いたのは相互メンション関係のみであり，言語情報は用いていない．また，コミュニティを抽出した後の特徴語群の生成には，ユーザの自己紹介文を用いている．ここで，各コミュニティの特徴語群がある程度意味的に類似していることを示すことができれば，自己紹介文が類似したユーザが同じコミュニティに所属していることとなり，結果として適切なコミュニティが得られていると考えられる．

そこで，下記のような手順を用いて，各コミュニティのコミュニティ内類似度 $\overline{S}$ を用いて，一定の類似性を持つことを示す．コミュニティ内類似度とは，各コミュニティにおいて，特徴語どうしのすべてのペアの類似度を算出し，その平均値をとった値である．この値が大きい方が，特徴語どうしの意味的類似性が高いため，コミュニティとしてよくまとまっていると考えられる．

(1) あるバースト事例に関わったユーザ群から，提案手法によりユーザコミュニティを抽出し，特徴語群を得る．そのうちコミュニティを構成するユーザ数の上位 N_{com} コミュニティを取り出す．
(2) (1) と同じユーザ群から，N_{member} 名をランダムに抽出し，ユーザコミュニティを N_{com} 個作成する．提案手法と同様の手法で，ユーザコミュニティごとに自己紹介文から特徴語群を得る．
(3) (1)，(2) の各ユーザコミュニティ U について，下記の式によりコミュニティ内類似度を算出する．なお特徴語群は上位 50 語を用いることとする．

$$\overline{S(U)} = \frac{2}{(N-1)(N-2)} \cdot \sum_{i=1}^{N} \sum_{j=i+1}^{N} Sim_w(w_i, w_j) \quad (2)$$

$$U = \{w_1, w_2, \ldots, w_N\}$$

表 3 御嶽山事例におけるコミュニティ内類似度の統計量の比較
Table 3 Similarity of communities in Ontake data.

手法	平均値	中央値	最大値	最小値	分散
提案手法	0.262	0.249	0.639	0.042	0.010
ランダム抽出	0.114	0.110	0.285	0.040	0.001

なお，上記手順において，各単語 w は単語分散表現ベクトルを用いて表すものとし，単語間類似度 Sim_w には 2 つの単語ベクトルの Cosine 類似度を用いた．単語分散表現とは，近年自然言語処理でよく用いられる単語のベクトル表現であり，特に類似語抽出に高い精度を発揮することが知られている [21]．単語分散表現の学習手法としては word2vec，学習コーパスとしては，日本語 Twitter 10%データ 3 カ月分（2012 年 1 月 1 日～3 月 31 日）を利用した．

御嶽山の事例について，各提案手法により作成した各コミュニティのコミュニティ内類似度を**表 3** に示す．ここでは $N_{com} = 2{,}335$，つまりすべてのコミュニティを利用することとした．実際には，2,335 個のコミュニティのうち，ユーザ 100 名以上から構成されるコミュニティ数は 1,858 であった．これらのコミュニティサイズに含まれるユーザ

の平均値が 1,694.02 であったため，$N_{member} = 1{,}694$ とした．表 3 より，ランダム抽出したコミュニティ群よりも提案手法で抽出したコミュニティ群の方がコミュニティ内類似度が 2 倍以上大きいことが分かる．

これより，提案手法により，相互メンション情報から，自己紹介文が類似したユーザ群がまとまったコミュニティを抽出することができたといえるだろう．

4.4 トピックエントロピー

各トピックがどの程度多様性のある人々によって拡散されたかを，トピック内の情報エントロピーを用いて評価する．トピック内の情報を拡散したユーザが所属するコミュニティがどの程度多様かを情報量の概念から求めたものをトピックエントロピーと呼ぶ．トピックエントロピーが小さければ，当該トピックは一部のコミュニティユーザによって拡散されたことを意味し，拡散数が多くても一部のユーザにしか届いていない可能性が高い．一方，トピックエントロピーが大きければ，多様なコミュニティのメンバによって広く拡散されたことを意味する．

トピック i のトピックエントロピー H_i は，以下のように求める．

$$H_i = -\sum_{c \in C} P(c) \log P(c) \tag{3}$$

ただし，C はコミュニティ集合，c はコミュニティ，$P(c)$ はあるツイートがコミュニティ c のメンバによって行われた確率を示す．なお，どのコミュニティにも所属していないユーザは，1 人で 1 つのコミュニティに所属していると見なす．また，bot[*3]であると判断されたアカウントは bot コミュニティに所属していると判断する．

図 6 に，御嶽山噴火に関して，ツイート数の多かった上位 10 トピック（表 1）のトピックエントロピーを示す．ほとんどのトピックで，トピックエントロピーはおおむね 5 以上であったが，T10 はトピックエントロピーが 0.740 と極端に小さかった．これは，このトピックを拡散したユーザの多様性が少ないことを意味する．

そこで，T10 に含まれるツイートを確認したところ，「富士山が噴火する（**図 7**）」「噴火の予知をした人物がいる」「噴火には UFO が関わっている」といった非現実的な内容であり，明らかなデマといってもよいものであった．このようなツイートが一般に受け入れられるとは考えにくい．実際，このトピックの情報は合計で 7,000 回近く拡散されているにもかかわらず，その拡散に寄与したユーザは 1,206 アカウントであり，そのうち 767 アカウントが bot であった．したがって，拡散された数は多いものの，このような情報は社会一般に広まったとはいえない．一方，表 1 より，T1～T9 は災害情報やニュースなど，何らかの新たな情報や安否情報や被災者による報告など，災害支援につながると推測される情報が含まれている．これらの情報は拡散された数も多く，また T10 と比較して相対的に多様な人々に広がったと考えることができるだろう．

以上のように，御嶽山の噴火におけるバースト事例では，その上位トピックのほとんどが高いエントロピーを持ち，多様性のある人々によって拡散されていたことがうかがえる．これより，T1～T9 に含まれる新たな情報を含むツイートや災害支援に関するトピックについては，社会に広く受け入れられていたと推測される．一方，bot が中心となって拡散されたデマを含むトピック T10 も存在していたが，そのような情報は多くの一般にはほとんど拡散していなかったことが明らかになった．これより，デマなどを含むトピックについては，ソーシャルメディア上での投稿数は多いものの，社会に広く受け入れられていなかったと

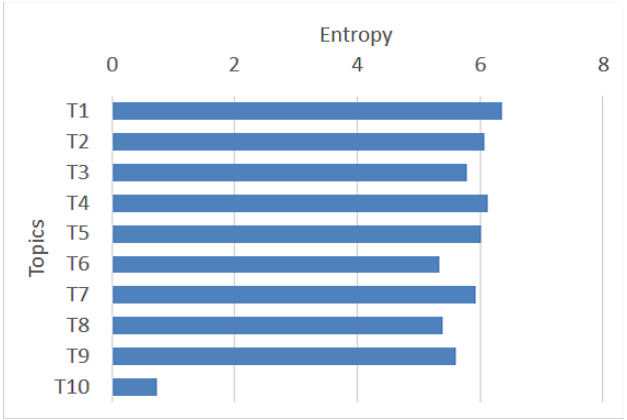

図 6 ツイート数上位 10 トピックのトピックエントロピー（御嶽山噴火）

Fig. 6 Tweet entropy of top10 topics (Ontake eruption).

図 7 T10 に含まれたツイートの例（御嶽山噴火）

Fig. 7 Tweet sample in T10 (Ontake eruption).

[*3] 自動投稿を行うアカウント

推測される．

本章では，御嶽山噴火の事例を詳細に分析し，提案手法の有用性について確認を行った．次章では，他のバースト事例についても分析を行い，個々のバースト現象がどのような性質を持っていたのかを明らかにする．

5. バースト現象におけるトピック分析

5.1 利用データ

本章では，Twitter 上で大きな話題となることが多い，災害，炎上，社会運動を事例として取り上げ，分析を行う．

対象となるバースト現象について，その種類，トピック名，および収集されたツイート数を**表 4** に示す．なお，各バースト現象の詳細については，付録 A.1 に示した．

5.2 トピックエントロピーの分布

各バースト事例ごとにトピックの抽出を行い，そのトピックエントロピーを求めた．まず，どの程度のエントロピーを持つトピックが抽出されたのかを確認するため，**図 8** に抽出されたトピックエントロピーの累積分布を示す．横軸にエントロピーを，縦軸に累積頻度を示している．ただし，規模が十分大きくないとエントロピーが小さくなる傾向にあることが分かっているため，総ツイート数が 1,000 以上のトピックのみを抽出の対象とした．

これより，ほとんどのバースト事例において，トピックエントロピーは 5 以上であることが分かる．トピックエントロピーが十分高ければ，拡散しているユーザが特定のコミュニティに偏っておらず，社会全体に拡散したトピックであるといえよう．一方で，いくつかのバースト事例では，トピックエントロピーが低い，すなわちユーザに偏りのあるトピックが存在していることが分かる．

御嶽山噴火の例で見たように，トピックエントロピーが極端に低いトピックには bot による投稿が多い場合もあるが，特定のコミュニティに所属するユーザが多く拡散している場合もトピックエントロピーは減少する．

そこで，具体的に各バースト現象についてトピックの分析を行う．

表 4 データセット
Table 4 Dataset.

種類	トピック名	ツイート数
災害	御嶽山噴火	1,097,091
災害	鬼怒川洪水	7,737,669
炎上	マクドナルド異物混入	959,792
炎上	オリンピックエンブレム問題	3,161,798
社会運動	アイスバケツチャレンジ	641,713

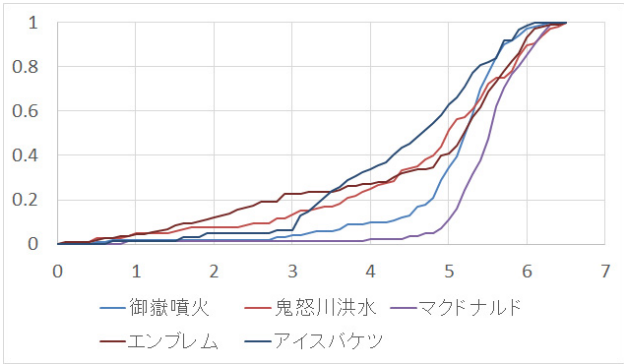

図 8 コミュニティごとのトピックエントロピーの累積分布
Fig. 8 Cumulative distribution of topic entropy.

5.3 トピックエントロピーを用いたトピックの分析

前節で見たとおり，いくつかのバースト事例においては，特定のコミュニティからのみ注目を受けたトピックが存在している．そこで，これらのトピックがどのようなものなのかを具体的に分析する．

図 9 に，各バースト事例におけるツイート数上位 10 のトピックについて，トピックエントロピーを求めたものを示す．個々に示しているのは，上位 10 件のトピックであるため，これらのバースト事例の中でも特に拡散されたトピックであるが，事例によってはこれらの中にもトピックエントロピーが低いものが混じっている．これらのトピックは特定のコミュニティのメンバによって拡散されたものであるため，どのようなコミュニティが興味を持ったのかは，分析に値するだろう．以下，トピックエントロピーの低いトピックに注目して，各バースト事例について分析を行う．

5.3.1 鬼怒川洪水

御嶽山の噴火と同じく災害事例である鬼怒川洪水は，御嶽噴火の事例と同じくほとんどのトピックでエントロピーが高い．災害という現象に関しては，基本的に社会一般に拡散するようなトピックが上位に来ることが多いようである．

ただし，8 番目のトピックに関してだけは，トピックエントロピーが低い．このトピックには，天気予報サイトのアカウントによる注意報に関するツイートおよび，災害に

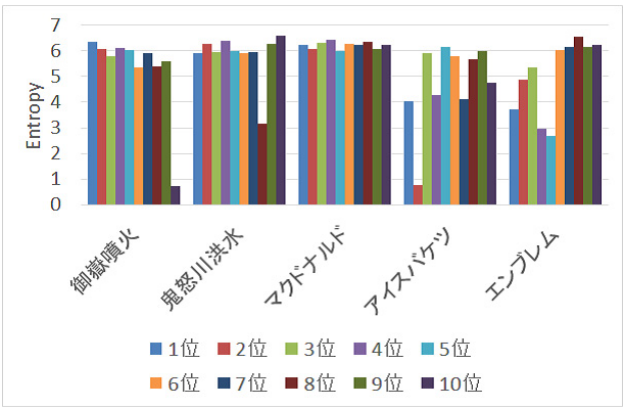

図 9 上位 10 トピックのトピックエントロピー
Fig. 9 Topic entropy.

関する注意喚起をするツイートが多く含まれていた．内容は一般的ではあるが，当該トピックを拡散したツイートの37.5%を写真コミュニティのメンバが行っていたことが分かった．注意報という重要な情報であるにもかかわらず，特定のコミュニティの中だけで拡散していたという事実は興味深い．これについては，さらに詳細な分析が必要である．

5.3.2 マクドナルド異物混入事件

マクドナルド異物混入事件については，大きく拡散したトピックのすべてでトピックエントロピーが高く，様々なコミュニティのメンバによって拡散されていたことが分かる．したがって，この問題については災害と同様，多様な人々が興味を持っていたと考えられる．

ただし，拡散数上位のトピックについてその中身を見ると，

(1) 異物混入を疑問視する内容
(2) マスコミの報道姿勢に対する批判
(3) 関連するネタ
(4) 異物混入に関する情報
(5) 無関係なツイート

となっていた．このうち，(1) はナゲットに歯が混入していたというニュースに対し「本人の歯ではないか」といった疑問や，「でっちあげではないか」という疑問など異物の混入そのものに対して疑問を呈したツイートが多く含まれていた．また，(2) では異物混入を報道するマスコミの報道の在り方に対する批判が中心であった．それ以外にも，異物混入を理由にマクドナルドを批判する内容を中心としたトピックは少ないことが明らかとなった．マクドナルドの異物混入事件は炎上事例であるが，企業批判以外の情報も拡散し，かつトピックエントロピー分析より，それは一部に閉じたものではなく社会全体に受け入れられていたことが示された．

5.3.3 東京オリンピックエンブレム問題

東京オリンピックエンブレム問題の事例では，上位5トピックのトピックエントロピーが低い．これらのトピックはいずれもエンブレム問題に対して批判的な意見およびブログやニュースサイトへのリンクである．最も拡散したトピック T1 における拡散数上位5コミュニティを**表 5** に示す．この問題については，政治コミュニティのメンバが大きく拡散していたことが分かる．これは，他の上位2〜5位までのトピックでも同様である．したがって，東京オリンピックエンブレム問題の批判的な意見は，政治コミュニティによって拡散されていたといえよう．

一方，上位6位のトピックは，デザイナによる肯定的な意見であり，7，9位はエンブレム発案者への攻撃を諫める意見であった[*4]．これより，批判的な情報の拡散は一部のコミュニティによって行われ，擁護的な情報がより幅広いコミュニティのユーザによって拡散されていたことが分かる．このような事実から，オリンピックエンブレム問題では集団極性化が発生し，エコーチャンバ現象が発生することで批判的な声が強くなった，あるいは強くなったようにみえた可能性がある．

[*4] 8位はキーワードにはマッチしたものの無関係なツイートによるトピックであり，10位は批判的な意見のまとめサイトへのリンクであった．

表 5 エンブレム問題上位1位トピックのコミュニティのラベル
Table 5 Community label for top 1 topic (Olympic emblem).

	拡散数	コミュニティラベル
C1	65,059	自由民主党（日本），日本共産党，日本
C2	2,759	Twitter，iPhone，映画音楽
C3	2,144	アニメ，CLANNAD（ゲーム），スワラ・プロ
C4	1,364	写真，心霊写真，証明写真
C5	1,245	大阪維新の会，維新の党，日本維新の会

表 6 アイスバケツチャレンジ上位1位トピックのコミュニティのラベル
Table 6 Community label for top 1 topic (Ice bucket challenge).

	拡散数	コミュニティラベル
C1	1,735	EXO，SHINee，ショー!K-POP の中心
C2	955	BIGBANG，YG エンタテインメントジャパン
C3	590	EXO，スーパージュニア，SHINee
C4	583	2PM，キミの声，チャンソン
C5	541	CNBLUE，チョン・ヨンファ

5.3.4 アイスバケツチャレンジ

アイスバケツチャレンジの事例では，上位2トピックのトピックエントロピーが低い．これらの2トピックの内容は，

(1) 韓国人芸能人によるチャレンジ報告（T1）
(2) まとめサイトへのリンク（T2）

であった．このうち，T1 を拡散したユーザのほとんどが韓国人芸能人のファンコミュニティに所属しており，これらのコミュニティに所属するユーザだけが拡散したトピックであったことが分かる（**表 6**）．一方，T2 では，このトピックに含まれるツイートを拡散したアカウントのうち70%が bot であることが確認され，これらの内容は大量に拡散されたものの，ほとんどが bot によって自動的になされたものであることが分かった．このような bot による投稿のほとんどは，ブログサイトなどへの誘導であるため，これらの情報が拡散したことが，社会一般にその情報が受け入れられたこととはならないため，注意が必要である．

また，ここで興味深いのが上位6位（T6），上位7位（T7）のトピックである．どちらも有名人によるチャレンジ報告であったが，T6 が人気バンドメンバによる報告，T7 がフィギュアスケートの選手らによる報告であった．それぞれトピックエントロピーは 5.789，4.135 となっており，フィギュアスケートの選手らによる報告の方がトピックエントロピーが低い．バンドメンバによる報告では，バンド

のファンコミュニティ以外に，高校生コミュニティや音楽コミュニティなどによる拡散が多く含まれているのに対し，フィギュアスケートの選手らによる報告では，フィギュアスケートファンコミュニティによる拡散がほとんどとなっている．同じ有名人の報告であっても，ファン層の違いが拡散される対象の違いに大きく関わっている点は興味深い．

6. おわりに

本論文では，ソーシャルメディア上でバースト現象が発生した際に，トピック分類，コミュニティ分類の手法を用いることで，誰がどのような意見を拡散していたのかを分析する手法を提案した．

トピック分類では，共通して興味を持つ情報に着目したネットワークを用いた分類手法を用いた．これによって，言語的手法である LDA を用いた分類よりも高い精度で情報の分類ができることを確認した．また，ユーザの分類ではコミュニケーション行動によるコミュニティ分類を行った．ここに，コミュニティの特徴語を用いて自動的にラベル付けをする手法を提案し，適切なラベル付けが行われていることを確認した．これらの手法を組み合わせ，各トピックにおいてどのようなユーザが拡散を行っているのかを明らかにするとともに，その多様性を情報エントロピーを用いて評価する手法を提案した．

本手法によって，ソーシャルメディア上で発生したバースト現象について，より詳細に分析することが可能となった．本論文で分析したいくつかのバースト事例の中には，広く拡散した情報が一般に思われている内容とは異なるものであったものや，bot による拡散が多いものなどが確認された．また，集団極化のようなソーシャルメディア特有の問題が発生している事例も確認された．集団極化はサイバーカスケード [22] を発生する要因になるとも考えられ，不適切なバーストの原因となりうる．計算社会科学の視点から，このような社会的に問題のある現象を早期に発見，防止する手法を確立するために，本提案手法は有用であると考えられる．

本提案手法は，必ずしも新しい手法ではなく，基本的には既存の手法の組合せによるものである．しかしながら，それによってバースト現象の詳細を分析することが可能であることを示した点が本論文における最も大きな貢献である．

今後の課題としては以下のようなものがあげられる．まず，現在用いている手法では，各情報は 1 つのトピックに，各ユーザは 1 つのコミュニティにしか所属していない．しかしながら，実際には複数のトピックにまたがるような情報や，複数のコミュニティに所属するユーザも存在するはずである．そこで，ソフトクラスタリングの技術を応用することで，このような問題に対応する手法を確立する必要がある．

また，バースト現象発生時には，bot によるツイートが多く見られる．bot による情報の拡散は，社会におけるバースト事例のとらえられ方を理解するうえでノイズとなるため，bot を的確に見つけ出し，分析対象から排除する必要がある．

最後に，本手法を用いて計算社会科学の視点から様々なバースト現象を分析し，その発生要因，防止策，あるいはそこから導き出される人間の本質を明らかにしていくことも大きな課題の 1 つである．

謝辞 本研究は日本学術振興会課題設定による先導的人文・社会科学研究推進事業「リスク社会におけるメディアの発達と公共性の構造転換」プロジェクトの一部として行われた．

参考文献

[1] Kleinberg, J.: Bursty and hierarchical structure in streams, *Data Mining and Knowledge Discovery*, Vol.7, No.4, pp.373–397 (2003).
[2] Toriumi, F., Sakaki, T., Shinoda, K., Kazama, K., Kurihara, S. and Noda, I.: Information Sharing on Twitter During the 2011 Catastrophic Earthquake, *Proc. 22nd International Conference on World Wide Web, WWW '13 Companion*, pp.1025–1028 (2013).
[3] García-Silva, A., Kang, J.-H., Lerman, K. and Corcho, O.: Characterising Emergent Semantics in Twitter Lists, *Proc. 9th International Conference on The Semantic Web: Research and Applications, ESWC '12*, pp.530–544, Springer-Verlag (2012).
[4] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X.: Comparing Twitter and Traditional Media Using Topic Models, *Proc. 33rd European Conference on Advances in Information Retrieval, ECIR '11*, pp.338–349, Springer-Verlag (2011).
[5] Ramage, D., Dumais, S. and Liebling, D.: Characterizing Microblogs with Topic Models, *Proc. ICWSM 2010*, American Association for Artificial Intelligence (2010).
[6] O'Connor, B., Krieger, M. and Ahn, D.: TweetMotif: Exploratory search and topic summarization for Twitter, *Proc. AAAI Conference on Weblogs and Social* (2010).
[7] Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I.: *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment* (2010).
[8] Rosa, K.D., Shah, R., Lin, B., Gershman, A. and Frederking, R.: Topical Clustering of Tweets, *Proc. ACM SIGIR: SWSM* (2011).
[9] Davidov, D., Tsur, O. and Rappoport, A.: Enhanced Sentiment Learning Using Twitter Hashtags and Smileys, *Proc. 23rd International Conference on Computational Linguistics: Posters, COLING '10*, pp.241–249, Association for Computational Linguistics (2010).
[10] Baba, S., Toriumi, F., Sakaki, T., Shinoda, K., Kurihara, S., Kazama, K. and Noda, I.: Classification method for shared information on Twitter without text data, *Proc. 24th International Conference on World Wide Web*, pp.1173–1178, ACM (2015).
[11] Tang, L. and Liu, H.: Community detection and mining in social media, *Synthesis Lectures on Data Mining and Knowledge Discovery*, Vol.2, No.1, pp.1–137 (2010).
[12] 金川元信，大豆生田利章：ソーシャルネットワークにお

けるリンク構造を用いた重複クラスタリング手法の提案，*DEIM Forum 2011* (2011).
[13] Mano, Y. and Aoyama, T.: USER CLUSTERING IN MINI-BLOG WITH FAVORITES OF USERS, 日本高専学会誌：Journal of the Japan Association for College of Technology, Vol.15, No.3, pp.43–46 (2010).
[14] 青島傳隼，福田直樹，横山昌平，石川 博：マイクロブログを対象とした制約付きクラスタリングの実現，第2回データ工学と情報マネジメントに関するフォーラム（DEIM2010）論文集 (2010).
[15] 松尾 豊，友部博教，橋田浩一，中島秀之，石塚 満：Web上の情報からの人間関係ネットワークの抽出，人工知能学会論文誌，Vol.20, pp.46–56 (2005).
[16] Clauset, A., Newman, M.E. and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol.70, No.6, 066111 (2004).
[17] Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E.: Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, Vol.2008, No.10 (2008).
[18] 篠田孝祐，榊 剛史，鳥海不二夫，風間一洋，栗原 聡，野田五十樹，松尾 豊：東日本大震災時における Twitter の活用状況とコミュニケーション構造の分析，知能と情報，Vol.25, No.1, pp.598–608 (2013).
[19] Arnaboldi, V., Conti, M., Passarella, A. and Pezzoni, F.: Ego networks in Twitter: An experimental analysis, *Proc. IEEE INFOCOM*, pp.3459–3464 (2013).
[20] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, Vol.3, pp.993–1022 (2003).
[21] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, pp.3111–3119 (2013).
[22] Sunstein, C.R.: *Republic.com 2.0*, Princeton University Press (2009).

付 録

A.1 利用データ

A.1.1 御嶽山噴火

取得キーワード

噴火 御嶽山 御岳山 火山 山小屋

データ概要

2014年9月27日に発生した御嶽山の噴火．登山者ら58名が死亡した，日本における戦後最悪の火山災害であった．

データ取得期間

2014/9/20–2014/10/9

A.1.2 鬼怒川洪水

取得キーワード

ゲリラ豪雨 ハザードマップ マグニチュード 仮設住宅 余震 停電 台風 噴火 噴石 地震 大雨 大雪 寒冷前線 帰宅困難 揺れ 断層 注意報 津波 洪水 活断層 浸水 減災 溶岩 火山 火砕流 災害 突風 竜巻 被災 警報 豪雪 避難 防災 集中豪雨 雪崩 震度 震源 震災

データ概要

2015年9月9日に発生した，台風18号にともなう豪雨災害．10日に茨城県常総市三坂町で堤防1カ所が決壊，11日には宮城県大崎市で鳴瀬川水系の渋井川の堤防が決壊した．

なお，キーワードが多いのは，災害一般として取得していたデータを利用しているためである．

データ取得期間

2015/9/8–2015/9/18

A.1.3 マクドナルド異物混入

取得キーワード

マクドナルド マック マクド

データ概要

2014年12月から2015年1月にかけて相次いでマクドナルドの商品から異物が発見された事件．

データ取得期間

2014/12/18–2015/1/14

A.1.4 オリンピックエンブレム問題

取得キーワード

エンブレム 佐野研二郎

データ概要

佐野研二郎氏が作成した2020年東京オリンピックのエンブレムが，模倣ではないかという疑いがもたれ，最終的にエンブレムが撤回された事件．

データ取得期間

2015/7/26–2016/9/21

A.1.5 アイスバケツチャレンジ

取得キーワード

ALS アイスバケツ アイスバケット Ice bucket 氷水

データ概要

2014年にアメリカで始まった，バケツに入った氷水を頭からかぶるかアメリカ ALS 協会（英語版）に寄付をするかを選ぶ運動．多くの著名人が氷水を被るとともに寄付を行い話題となった一方で，批判的な意見も相次いだ．

データ取得期間

2014/8/1–2014/9/2

鳥海 不二夫 （正会員）

2004 年東京工業大学大学院理工学研究科機械制御システム工学専攻博士課程修了．同年名古屋大学情報科学研究科助手．2007 年同助教．2012 年東京大学大学院工学系研究科准教授．エージェントベースシミュレーション，ソーシャルメディア，計算社会科学，ゲームにおける AI 等の研究に従事．人工知能学会，電子情報通信学会，日本社会情報学会各会員．博士（工学）．

榊 剛史

2004 年東京大学工学部電子情報工学科卒業．2006 年同大学大学院修士課程修了．電力会社通信部門での勤務を経て 2009 年同大学院博士課程入学．2014 年同博士課程修了．博士（工学）．東京大学での特任研究員を経て，2015 年より現職ならびに東京大学客員研究員．専門は，Web マイニング，自然言語処理，計算社会科学．