情報処理学会論文誌　データベース　Vol.11　No.3　35–46 (Oct. 2018)

バースト現象を考慮したハッシュタグクラスタリング

福山 怜史[1,a)]　若林 啓[2,b)]

受付日 2018年3月19日, 採録日 2018年7月4日

概要：Twitter においてバースト現象が確認されたハッシュタグを収集することは，現実世界で流行している話題を抽出するうえで重要なアプローチである．しかしハッシュタグには表記揺れや様々な抽象度を持つタグが混在する性質があるため，同じ話題を指すハッシュタグが複数存在していたり，バーストしていないタグでもバーストタグと同じ話題を示す可能性がある．この問題に対する 1 つの解決策としてハッシュタグのクラスタリングが考えられるが，一度にすべてのハッシュタグのクラスタリングを行う処理は計算コストが高いため，効率的な手法が必要となる．本研究では，以上の問題を効率的に解決することを目的として，バーストタグのみクラスタリングを行い，生成したクラスタに非バーストタグを割り当てる手法を提案する．これによりクラスタリング処理を行うハッシュタグはバーストタグだけになるため，クラスタリングに要する実行時間が短縮できる．提案手法では，3σ 法によってハッシュタグのバースト検出を行い，バーストタグを spherical k-means 法によってクラスタリングし，生成したクラスタの中から最も重心の近いクラスタに非バーストタグを割り当てる．実験により，一度にすべてのハッシュタグをクラスタリングする手法と比較して，話題のまとまりの良さを保ちながら，実行時間が短縮できることを示す．

キーワード：Twitter，ハッシュタグ，バースト検出，クラスタリング

Hashtags Clustering for Discovering Bursty Topics

Satoshi Fukuyama[1,a)]　Kei Wakabayashi[2,b)]

Received: March 19, 2018, Accepted: July 4, 2018

Abstract: Collecting bursty hashtags in Twitter is a promising approach to discover popular topics in the world. However, a set of hashtags, which can be a mixture of bursty and non-bursty ones, potentially represents the same topic because the hashtags are user-generated labels that have inconsistent spellings and granularities. Therefore, we need to apply the method to aggregate hashtags that indicate the same topic. One of the method is clustering, but clustering over all hashtags in Twitter is very expensive regarding the computational cost. In order to solve this problem, we propose an efficient bursty hashtag clustering method that consists of two procedures; (1) a clustering of bursty hashtags, and (2) an assignment of each non-bursty hashtag to the nearest cluster. Since the clustering step processes only the bursty hashtags, the proposed method can reduce the total execution time compared with the method that conducts a clustering of all hashtags at the same time. We employ 3σ method and spherical k-means for the bursty hashtag detection and clustering. Experiments with human judgments suggest that our method keeps coherent tags and saves more times in comparison with the all hashtags clustering.

Keywords: twitter, hashtag, burst detection, clustering

[1] 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[a)] s1721691@s.tsukuba.ac.jp
[b)] kwakaba@slis.tsukuba.ac.jp

1. はじめに

Twitter は，現実世界で起こった事象に対してユーザがリアルタイムにツイートを投稿する性質から，現実世界を知覚するセンサとしての利用が期待されている．たとえば，2011 年 3 月 11 日に発生した東日本大震災では東京

都において地震発生から 1 時間以内に毎分 1,200 件以上のツイートが投稿されたことが報告されている [1]．近年では，Twitter からの評判情報抽出 [2] や病気の流行予測 [3] といった手法の有効性が確認されており，Twitter ユーザが現実世界の事象に対して敏感に反応することが分かる．

このような特徴から，Twitter 上の投稿の傾向を分析することで，現実世界で起きた出来事や流行している話題を抽出する手法が研究されている [4], [5], [6], [7]．これらの手法では，局所的な時間で話題の出現頻度が急激に増加する “バースト現象” を検出することで，出来事や流行の抽出を行う．このように特定の短い期間に注目を集めた話題を抽出し，その話題に関連するツイートを網羅的に収集することは，現実世界の事象の調査や，特定の人物や商品の評判分析などにおいて有用である．

Twitter では，ユーザが特定の話題についての言及していることを明示するために，ハッシュタグと呼ばれる「#」を付けた文字列をツイートに含めることが一般的である．このため，ハッシュタグの投稿頻度の時系列変化から，頻度が急激に増加するようなバースト現象を検出することで，Twitter において流行している話題を抽出できる．本稿では，このように投稿頻度が急激に増加したハッシュタグを，当該期間における “バーストタグ” と呼ぶ．またバーストタグ以外のハッシュタグを “非バーストタグ” と呼ぶ．

しかし，ハッシュタグはユーザが自由に作成できるため，表記揺れが存在したり，様々な抽象度のハッシュタグが Twitter 上に混在したりしている [8]．このことに起因して，単純にバーストタグを列挙して流行した話題を抽出する方法には，2 つの問題がある．1 つは，バーストタグの集合の中には，同じ話題を指し示しているものが重複して含まれており，話題の抽出手法として冗長な出力になるという問題である．もう 1 つは，非バーストタグの中にも，バーストタグと同じ話題を指すものが存在する可能性があり，関連ツイートの網羅性が損なわれるという問題である．たとえば，オリンピックに関係する「#日本代表」というハッシュタグがバーストしているときに，関連した非バーストタグ（特定の選手に注目した「#本田圭佑」など）を抽出することは，当該の話題の全容を把握するうえで重要である．これらの観点を整理した例を**図 1** に示す．図 1 では，発生したハッシュタグに対して，3 種類の話題が流行していると推定できる．

異なるハッシュタグが同じ話題を指し示しているかどうかは，当該のハッシュタグと共起する単語の類似性に基づいて判別できると考えられる．Tsur ら [9]，井上ら [10] は，共起する単語を特徴量としてハッシュタグを k-means 法を用いたクラスタリングを行うことによって，表記揺れや階層関係を吸収し，事象ごとのクラスタを構築する手法を提案した．この手法を素直に利用して前述の問題を解決するためには，前もって当該期間に発生したすべてのハッシュタグのクラスタリングを行ったうえで，バーストタグを含むクラスタを出力する必要がある．しかし k-means 法では，クラスタ数とハッシュタグ数の積に比例して実行時間が増加するため，すべてのハッシュタグを話題ごとにクラスタリングするアプローチは計算コストが大きい．本研究の目的を達成するためには，バーストタグと関連した話題のみをクラスタリングできれば十分であるため，より効率的な手法が検討できる可能性がある．

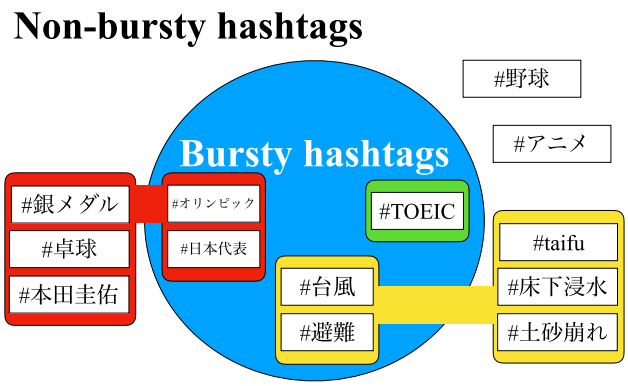


図 1　ハッシュタグと Twitter で流行している話題の関係
Fig. 1　The relationship between hashtags and bursty topics on Twitter.

本研究[*1]では，バーストタグのみを用いてクラスタリングを行い，非バーストタグを生成されたクラスタに割り当てる手法を提案する．これにより，クラスタリング処理ではバーストタグのみを考慮すればよいため，クラスタリングにかかる実行時間を大幅に短縮できる．本稿では，提案手法を用いることで，先行研究を直接適用してすべてのハッシュタグをクラスタリングする場合と比較して，話題のまとまりの良さを保ちながら，実行時間が短縮できることを示す．

2.　関連研究

本研究に関連して Twitter におけるバースト現象に関する研究とハッシュタグのクラスタリングに関する研究がある．まず Twitter におけるバースト現象に関する研究として水沼ら [4] の研究がある．水沼らは Twitter におけるバーストの特徴を分析しその特徴ごとにバーストの類型化を行っている．水沼らはツイートのバースト検出手法を選択するために，バースト現象を出現頻度の外れ値と見なし，外れ値検知手法である ROKU [12]・3σ 法・増山の検定 [13]・MAD 法 [14], [15] の比較を行っている．この比較では，1 度のバースト現象を複数のイベントと見なすことがないか，4 手法でバースト現象の検出漏れがどれぐらい少ないかを評価している．この結果 3σ 法が，バースト平均継続時間が最も長く，また最も検出漏れの低い，4 手法の中で最も理想的な手法であるとしている．また各手法に

[*1] 本稿は研究会論文 [11] を発展させ，手法および実験を追加したものである．

おいて，1 回のバースト検知で，ROKU ではエントロピー，3σ 法と増山の検定では平均および標準偏差，MAD 法では中央値を計算しており，実行時間のオーダは時系列データの窓幅は等しく，各手法で大きな違いはない．以上の結果から，水沼らが最も理想的な手法と見なす 3σ 法をバーストタグの検出手法に用いる．

次にハッシュタグのクラスタリングに関する研究として，Tsur ら [9] の提案する事象ごとにハッシュタグをクラスタリングする手法がある．この研究ではハッシュタグごとに当該のハッシュタグを含むツイートをすべて結合した擬似文書を作成し，その擬似文書をクラスタリングする手法を提案している．Tsur らは特徴ベクトルの作成に TF-IDF ベクトルやハッシュタグの共起ベクトルを用いており，クラスタリング手法には k-means 法を用いている．k-means 法とは，クラスタ重心の計算とデータのクラスタ割り当てを，割当てが変化しなくなるまで繰り返す手法であり，クラスタ重心の計算・クラスタの割当てにはユークリッド距離が用いられることが多い．Tsur らは出現頻度の高いハッシュタグ 10,000 種類に対してクラスタ数 1,000 で k-means 法を適用した．この結果，獲得したクラスタには潜在的なトピックを持つもののほかに表記揺れを吸収するタグクラスタが得られたことを報告している．また，井上ら [10] は，Tsur らが英語で行ったハッシュタグのクラスタリングが，日本語でも適用できることの検証を行った．この結果，クラスタ数を全体のハッシュタグ数に対して，60–70%の割合に設定することで表記揺れを吸収するタグクラスタが得られることを報告している．

しかし，TF-IDF ベクトルは L_2 ノルムによって正規化され超球面上に分布しているため，ユークリッド距離を用いた k-means 法ではクラスタ重心の位置が超球面上に位置せず，データ点のクラスタを正確に確率分布でモデル化できていないといえる．この問題を解決するために，Tsur と井上らの手法における k-means 法を Dhillon らが考案した spherical k-means 法 [16] に変更する方法が考えられる．spherical k-means 法は，一般的な k-means 法に対して，クラスタ重心の計算の際に重心ベクトルを L_2 ノルムによって正規化する処理を加え，データ点とクラスタ重心との距離にコサイン類似度を考慮したクラスタリング手法である．自然言語の TF-IDF ベクトルの場合はベクトルの大きさに意味がなく，そのベクトルの方向にのみ意味があるため，spherical k-means 法はユークリッド距離を用いた k-means 法と比較して，クラスタをより正確に確率分布でモデル化できると考えられる．本研究ではハッシュタグごとにツイートをすべて結合した擬似文書から TF-IDF ベクトルを作成し，spherical k-means 法によってハッシュタグをクラスタリングする．

3. 提案手法

本研究では，バーストタグの検出，バーストタグのクラスタリング，生成したクラスタへの非バーストタグの割当てを行う手法を提案する．バーストタグの検出では，3σ 法と分位数による閾値を超えた出現頻度を持つハッシュタグをバーストタグと見なす．次に検出したバーストタグの含まれたツイートから TF-IDF ベクトルを作成し，spherical k-means 法によってクラスタリングする．そして非バーストタグも同様に TF-IDF ベクトルを作成し，バーストタグのクラスタに割り当てる．この際，図 1 での「#野球」や「#アニメ」のように，非バーストタグが話題的に関連するバーストタグのクラスタが存在しない場合がある．このような非バーストタグは，関連性の低いいずれかのクラスタに割り当てられてしまい，クラスタのまとまりを悪くするため，関連度の低い非バーストタグのフィルタリング処理を行う．

3.1 ハッシュタグにおけるバースト現象の検出方法

本研究では，ハッシュタグのバースト検出に 3σ 法を用いる．3σ 法とは，外れ値検出手法の 1 つで，あるデータ x_t が，以下の条件を満たすとき，x_t を外れ値とする手法である．

$$x_t > \mu + 3\sigma \tag{1}$$

μ は時系列データ X における平均値，σ は標準偏差である．この手法をハッシュタグのバースト検出で行う場合，過去数時刻分のハッシュタグの出現頻度が与えられたデータ集合から，平均と標準偏差を計算し，平均に標準偏差の 3 倍を加えた値をバースト閾値とする．そして閾値を超える出現頻度が観測された場合，バーストしたと見なす．

ここで 3σ 法では，出現頻度が少ない場合，バースト検出がされやすいという問題がある．たとえば，$X = (2, 4, 0, 2, 6)$ の時系列データが与えられた場合，バースト閾値が 10 となるが，出現数 10 をバーストしたと見なすことは直感的に少なすぎると考えられる．そこで本研究では，同じ時刻に発生しているすべてのハッシュタグの出現頻度を考慮し，バースト検出における出現頻度の下限を設ける．ここで各ハッシュタグの出現頻度には偏りが存在することが分かっているため，バースト検出における出現頻度の下限をこの出現頻度の偏りに影響されにくい分位数によって決定する．

3.2 ハッシュタグのクラスタリング

バーストタグの集合を H' とする．ここでは，H' の要素をクラスタリングすることで，同じ話題を指し示すハッシュタグのクラスタを得ることを目指す．

クラスタリングに用いるハッシュタグの特徴量として，

共起する単語の TF-IDF ベクトルを用いる．ハッシュタグ $h \in H'$ が出現するツイートの集合を D_h とする．ツイート $d_h \in D_h$ に語彙 w_i が出現する頻度を $tf(w_i, d_h)$ と表すと，ハッシュタグ h と共起する単語の頻度は以下のように定義される．

$$tf(w_i, h) = \sum_{d_h \in D_h} tf(w_i, d_h) \tag{2}$$

また，語彙 w_i の文書頻度を以下のように定義する．

$$df(w_i) = \sum_{h \in H'} U(tf(w_i, h) - 1) \tag{3}$$

ここで，$U(x)$ は $x \geq 0$ のとき $U(x) = 1$，$x < 0$ のとき $U(x) = 0$ となる単位ステップ関数である．これを用いて，語彙 w_i の逆文書頻度 $idf(w_i)$ を以下のように定義する．

$$idf(w_i) = \log \frac{1 + |H'|}{1 + df(w_i)} + 1 \tag{4}$$

語彙の集合を W としたとき，ハッシュタグ h の正規化されていない TF-IDF ベクトル $\boldsymbol{v}'_h = (v_{h,1}, \ldots, v_{h,|W|})$ は，各要素が以下のように表される $|W|$ 次元のベクトルとして定義される．

$$v_{h,i} = tf(w_i, h) \cdot idf(w_i) \tag{5}$$

ハッシュタグ h の特徴量は，$|W|$ 次元の正規化された TF-IDF ベクトル $\boldsymbol{v}_h$ として定義する．

$$\boldsymbol{v}_h = \frac{\boldsymbol{v}'_h}{\|\boldsymbol{v}'_h\|} \tag{6}$$

以上の方法ですべてのハッシュタグにおけるベクトルを作成し，spherical k-means 法を用いてクラスタリングを行う．すべてのハッシュタグの集合 $\boldsymbol{v}_h$ を，spherical k-means 法によってクラスタリングを行う．

3.3　非バーストタグのクラスタへの割り当て

spherical k-means 法ではデータ点が超球面上に分布するため，クラスタ重心とデータ点とのコサイン類似度を θ としたとき，これら 2 点の距離は $1-\theta$ で計算できる．よってクラスタに関連のないタグをフィルタリングする場合，直感的にはコサイン類似度に対して適当な閾値を設けてフィルタリングを行うことが考えられるが，コサイン類似度は多次元空間においては次元の呪いの影響を受けるため，同一データでも次元数が増えるに従ってコサイン類似度の分布が変化する．このため，データの次元数によって異なるコサイン類似度の閾値を設ける手続きが必要となってしまう．この場合，次元数はデータセットに出現する単語数を表すため，すなわち単語数に応じてコサイン類似度の閾値を変更しなければならない．この問題を防ぐため，本研究ではクラスタ重心とデータ点とのコサイン類似度を確率分布によってモデル化し，累積確率を閾値に設定することで，単語数に応じてコサイン類似度を変更せずにフィルタリングを行う．

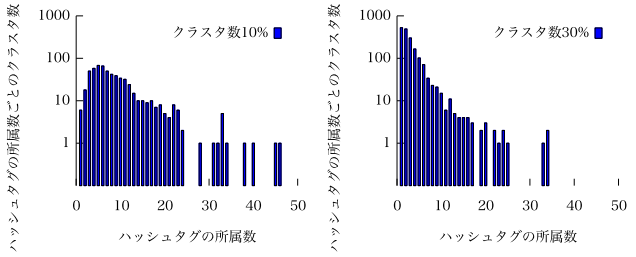


図 2　クラスタごとに割り当てられるハッシュタグ数の偏り．割り当てられるハッシュタグ数が少ないクラスタが多く分布している．

Fig. 2　The number of hashtags allocated in each cluster. Many clusters have a little number of hashtags to be allocated.

この spherical k-means 法に対して，データの集中度を考慮した手法である混合 von Mises-Fisher 分布では，データ点とクラスタ重心の距離はカイ二乗分布でモデル化できると井出ら [17] は報告している．しかし，このモデル化では $1-\theta \ll 1$ であることを仮定しているため，クラスタ重心に非常に近い領域しか適切にモデル化できないと考えられる．実際に評価実験と同一のデータを用いた予備実験[*2]で井出らの手法を用いたが，ハッシュタグとクラスタ重心の距離の分布と井出らの手法によって推定したパラメータのカイ二乗分布を比較したところ，あてはまりが悪いことが分かった．このため本研究では，計算が容易であり，定義域が spherical k-means 法の距離の範囲 $[0, 1]$ と等しいベータ分布を用いる．

本研究ではハッシュタグとクラスタ重心のコサイン類似度 θ_h をベータ分布でモデル化し，モーメント法によってベータ分布のパラメータを点推定する．ここで，データ点とクラスタ重心の距離は本来 $1-\theta$ によって計算されるが，行われているフィルタリング処理はコサイン類似度でも等価なため，距離の代わりにコサイン類似度を用いる．そして，この分布において発生確率の低いコサイン類似度を持つ非バーストタグをフィルタリングすることによって，関連度の低いハッシュタグのクラスタへの割当てを防ぐ．ここでクラスタに割り当てられるバーストタグの数はクラスタごとに偏りがあることが評価実験と同一のデータを用いた予備実験において**図 2** に確認されており，割当てタグ数の少ないクラスタでは，クラスタ内で観測されたデータに過適合したベータ分布となる可能性がある．この問題に対応するため，本研究では，他のクラスタのハッシュタグと重心とのコサイン類似度を考慮し，すべてのクラスタで単一のフィルタリングを行う．提案手法では，各クラスタにおけるハッシュタグと重心とのコサイン類似度を計算し，

[*2] 評価実験と同一のデータを用いているため，過適合を引き起こしている可能性があるが，本研究ではそのハイパーパラメータの性質は結果に大きく依存していないと仮定している．以降の予備実験でも同様である．

それらすべてのコサイン類似度から平均と分散を計算し，ベータ分布のパラメータの点推定に用いる．

$\Theta = \{\theta_h | h \in H'\}$ をバーストタグ h とその所属クラスタの重心とのコサイン類似度 θ_h の集合とし，モーメント法によって Θ の分布をベータ分布によって点推定する．ハッシュタグ h のコサイン類似度 θ_h およびパラメータ α と β を持つベータ分布の確率密度関数は，(7) によって計算される．

$$Be(\alpha, \beta) = \frac{\theta_h^{\alpha-1}(1-\theta_h)^{\beta-1}}{B(\alpha, \beta)} \tag{7}$$

$Be(\alpha, \beta)$ はベータ分布であり，$B(\alpha, \beta)$ はベータ関数である．そしてベータ分布のパラメータ α と β を (8) と (9) を用いて点推定する．

$$\alpha = \mu_\Theta^2(1-\mu_\Theta)\frac{1}{s_\Theta} - \mu_\Theta \tag{8}$$

$$\beta = \frac{\alpha}{\mu_\Theta} - \alpha \tag{9}$$

μ_Θ と s_Θ は，それぞれ Θ の平均と分散を表す．以上の式によって点推定されたベータ分布において，出現確率の下限を p，非バーストタグ $\hat{h}$ の重心とのコサイン類似度を $\theta_{\hat{h}}$ とすると累積分布関数 $F(\theta_{\hat{h}}) = p' > p$ となる $\hat{h}$ のみクラスタに割り当てる．

次に非バーストタグの TF-IDF による特徴量ベクトルを作成する．この際，3.2 節で生成したクラスタと同じ特徴量空間に写像する必要があるため，特徴量を抽出する語彙と IDF の値はバーストタグの特徴量と同じものに設定する．そして作成された各ハッシュタグの特徴量ベクトルから，各クラスタが持つ重心の位置までのコサイン類似度を計算し，最近傍のクラスタを選択する．この選択されたクラスタとのコサイン類似度が，計算された距離の閾値を超えていた場合，特徴量ベクトルが示すハッシュタグのクラスタと見なす．

4. 評価実験

本研究の提案する手法を評価するために，以下の点について確認を行った．

(1) 提案手法によって生成したハッシュタグのクラスタが分かりやすくまとまっているか
(2) 生成したクラスタに対して，非バーストタグが適切に割り当てられているか
(3) ベースラインと比較した場合，クラスタリング処理の実行時間は短縮されているか

(1) では，クラスタのまとまりの良さを評価するために，Chang ら [18] の提案した Word intrusion を拡張した Tag intrusion を考案し，これによってユーザ検証を行った．(2) では，提案手法とベースラインで非バーストタグの割り当てに対して，Tag intrusion の結果に違いがあったか確認を行った．(3) では，ベースライン手法と比較して，実行時間がどの程度短縮されたかについて確認した．

4.1 ベースラインについて

本研究の目的は流行している話題に関係するハッシュタグのクラスタの獲得であるため，ベースラインでは一度にすべてのハッシュタグをクラスタリングを行う井上ら [10] の手法にハッシュタグのバースト検出処理を加えたものとする．クラスタリングを行うハッシュタグは 4.2 節で示したバーストタグと非バーストタグを組み合わせたハッシュタグ集合を用いる．またクラスタリング手法は spherical k-means で行う．

4.2 実験データおよび環境

実験対象とするハッシュタグは，2012 年 8 月 1 日から 2012 年 8 月 7 日の間に存在したハッシュタグ 965,306 種類である．はじめにハッシュタグのバースト検出における設定について述べる．本研究では，時系列の単位を 1 日，バースト閾値を計算する時系列データの窓幅を 30 日分とする．時系列の単位を 1 日とする理由は，ツイート全体の投稿数が増加しやすい時間帯では誤ってバースト検出される可能性があるため，この問題を考慮する現状の制約として設けた．時系列の窓幅の設定では，その幅の長さによって獲得できるバースト現象は長期的なバーストから短期的なバーストが変化することから，本研究では話題の流行り廃れが長くても 1 カ月以内であるバースト現象を対象とし窓幅を 30 日に設定した．以上より，3σ 法による閾値を当日 t を含まない過去 30 日分の投稿数 $X = (x_{t-30}, \ldots, x_{t-1})$ から計算する．この場合 2012 年 8 月 1 日の閾値は，2012 年 7 月 2 日から 2012 年 7 月 31 日の各日のツイート数の集合から計算される．

またバーストタグの出現頻度の下限に応じたバースト検出されるタグの数を確認するため評価実験と同一のデータを用いた予備実験を行った．この予備実験では出現頻度の候補の下限ごとに 3σ 法によってバーストタグの検出を行う．出現頻度の下限の候補として，中央値・第 1 四分位数・第 1 五分位数・第 1 十分位数としたときの当該期間における出現頻度の下限および検出されたバーストタグの数の平均値と標準偏差を**表 1** に示す．この結果より出現頻度の下

表 1　各分位数を出現頻度の下限とした場合の検出バーストタグ数
Table 1　The number of detected hashtags according to the lower limit of occurrence frequency We selected the lower limit of the frequency of occurrence from each quantile.

下限の候補	出現頻度の下限（平均 ± 標準偏差）	検出タグ数
第 1 十分位数	10.3 ± 0.9	35,389
第 1 五分位数	57.4 ± 5.4	8,420
第 1 四分位数	111.0 ± 10.9	4,776
中央値	1,160.6 ± 152.3	494

限は低いほど検出されるバーストタグ数が多く高いほど少なくなることから，下限の設定が低すぎる場合ではわずかな出現頻度の増減によってバースト検出されてしまい，設定が高すぎる場合では過剰にバーストタグをとり除いてしまう可能性があることが分かった．以上の結果をふまえた出現頻度の下限として，値が高すぎず，低すぎないと考えられる第 1 五分位数を設定した

本研究ではバースト検出されたタグの中から以下の制約を満たすバーストタグ 6,033 種類・非バーストタグ 18,136 種類，計 24,169 種類に対して実験を行った．

(1) 各ハッシュタグが与えられたツイートに出現する単語の総数がタグの出現数以上である．
(2) 当該期間における非バーストタグの総出現数は 2012 年 8 月 1 日から 8 月 7 日におけるバーストタグの出現頻度の下限の平均値を超える．

1 つ目の制約では，ハッシュタグの一部には本文に出現する語の数が乏しいものが存在しこれらのハッシュタグから特定の話題を獲得することは困難であることから，少なくとも 1 ツイートに 1 単語以上が出現しているハッシュタグのみをクラスタリングするために設けた．2 つ目の制約では，提案手法および先行研究の手法において膨大な種類の非バーストタグすべてをクラスタに割り当てることが困難であることから，クラスタ割り当てに用いるハッシュタグを限定する必要があるため設けた．3σ 法によるバースト現象の検出では，バースト時刻では局所的に出現頻度が増加しそれ以外の時刻では出現頻度が低いタグが検出されることから，当該期間におけるバーストタグの総出現数の下限は提案手法で採用したバーストタグの出現頻度の下限と近い値になる傾向があると考えられる．以上をふまえて本研究では，各日のバーストタグの出現頻度の下限の平均値を計算し，この値を非バーストタグの総出現数の下限として，これらの非バーストタグに限定する．バーストタグの出現頻度の下限は当該時刻の全ツイートの出現頻度の第 1 五分位数であり，2012 年 8 月 1 日から 2012 年 8 月 7 日における各日の出現頻度の下限からその平均値を計算する．

各ハッシュタグに用いるツイートはハッシュタグと同一期間に発生した 25,897,072 ツイートを用いた．実験環境は，OS が Ubuntu 16.04，CPU が Intel Xeon E5-2630（2.40 GHz）8 core 2 機であり，実装は Python で行った．形態素解析器は MeCab [19]，TF-IDF および spherical k-means 法はオープンソースの機械学習ライブラリの 1 つである scikit-learn を利用し，spherical k-means 法は k-means 法のモジュールに重心の正規化処理およびコサイン類似度による距離計算の処理を加えた．また spherical k-means 法の初期重心ベクトルは k-means++法 [20] によって確率的に選択した．

4.3　ハッシュタグの特徴量ベクトルの設定

本研究の検証では，特徴量ベクトルの次元数が膨大になることを防ぎかつ特定の話題を表す傾向の強い品詞に限定することを目的として，TF-IDF で使用するコーパス内の語彙に以下の制約を設ける．

(1) $df(w) \geq 10$ であるような語彙
(2) 語彙の品詞は固有名詞・普通名詞・サ変接続の名詞のみ
(3) 語彙の文字列長は，漢字は 1 字以上・ひらがな・カタカナ・数字・記号では 2 字以上
(4) リツイートを示す「RT」・URL・リプライを示す「@ユーザ名」の文字列は無視

この制約によって得られた語を特徴語として，それぞれのハッシュタグにおける TF-IDF の特徴量ベクトルを作成し，spherical k-means 法によってクラスタリングを行う．このときに，語彙の制約によって零ベクトルとなるハッシュタグが発生するため，本研究ではこのようなハッシュタグはクラスタリングに考慮しない．

4.4　各種パラメータの設定

クラスタ数の設定では Tsur らが任意のトピックを示すタグクラスタの獲得を行うために設定しているものと同様のハッシュタグ数に対する 10%と新たに 30%を加えたものでクラスタリングを行う．30%のクラスタに関しては，井上らの手法がクラスタ数がハッシュタグ数に対する 60%であり，Tsur らと井上らの中間の大きさのクラスタ数として採用している．提案手法では，零ベクトルでないバーストタグが 6,028 種類であり，クラスタ数はバーストタグ全体の 10%の 603 と 30%の 1,808 とした．そして提案手法によって生成したクラスタに対して，零ベクトルでない非バーストタグ 18,129 種類の割り当てを行った．この結果，フィルタリングされずに割り当てられたタグは，クラスタ数が 10%の場合では 8,197 種類が，クラスタ数が 30%の場合では 3,829 種類が割り当てられた．またベースラインによるクラスタリング結果は，提案手法と同様の語彙の制約のうえで，一バーストタグと非バーストタグを組み合わせたハッシュタグ集合に対してクラスタリングを行う．この場合，零ベクトルでないハッシュタグが 24,161 種類であり，クラスタ数は提案手法のクラスタ数と同じ割合である 10%の 2,416 と 30%の 7,248 でクラスタリングを行った．ここで提案手法とベースラインでハッシュタグの総数が 24,157 と 24,161 でありそれぞれが異なる理由は，制約「$df(w) \geq 10$ であるような語彙」において，提案手法では満たされないが先行研究ではこの制約が満たされたハッシュタグが存在するためである．以降では “クラスタ数 n%” はクラスタリングするタグの種類数に対し n%に相当するクラスタ数を設定した手法を示す．

また出現確率の下限 p の決定のために p の候補（0.05，0.10，0.15）で評価実験と同一のデータを用いた予備実験

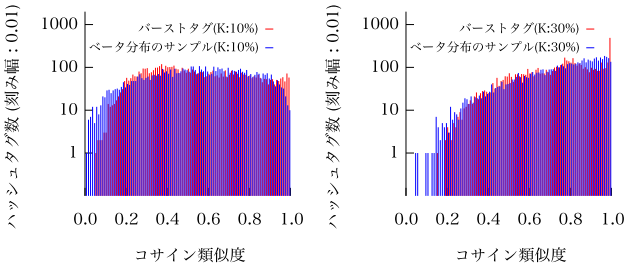


図 3　バーストタグとそのクラスタ重心とのコサイン類似度と点推定されたベータ分布の比較

Fig. 3　The comparison between the cosine similarity of the bursty tag to the cluster centroid and the estimated beta distribution.

を行った．この結果，クラスタに割り当てれらた非バーストタグのタグ名やタグに関連するツイートから一部の内容を目視して主観的に判断し，$p = 0.10$ を出現確率の下限として用いた．このときのバーストタグと所属クラスタの重心とのコサイン類似度および推定したベータ分布からのサンプルのヒストグラムを**図 3** に示す．図 3 に示すベータ分布のサンプル数はバーストタグと同じである．

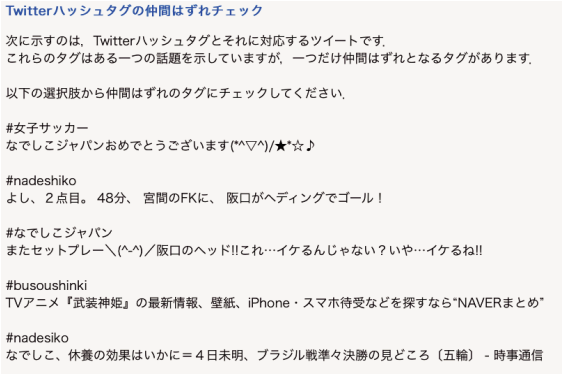
4.5　Tag intrusion によるユーザ検証

本研究では，4.2 節で述べたデータと 4.4 節で述べた設定を適用した提案手法から，ハッシュタグのクラスタを生成し，これらのクラスタのまとまりのよさを評価するためにユーザ検証を行う．この検証との親和性が高い検証手法として Chang ら [18] の Word intrusion がある．Word intrusion とは，トピックモデルにおいて人にとって分かりやすくまとまっているか定量的に評価するための方法である．この方法では，トピック内で出現確率が最も高い語を複数個と出現確率が最も低い語 1 つを選択肢を持つタスクから，これらの選択肢からユーザが仲間はずれだと考える語を回答してもらい，出現確率の低い語が選ばれた割合からそのトピックの人にとっての分かりやすさを定量的に評価する．本研究では Word intrusion におけるトピックをクラスタに，語をハッシュタグに置換した検証を行う．この際，ハッシュタグ自体はあくまでラベルであり，そのタグが示す話題は実際のツイート内容に含まれているため，ハッシュタグとそのハッシュタグが付与されたツイートを同時にユーザに提示する．この検証手法を，本研究では明示的に “Tag intrusion” とする．

本研究では，2017 年 12 月 6 日に Lancers [*3] で提案手法と井上らの手法を評価する Tag intrusion の実験を行った．設定したタスクの詳細を以下に示す．

(1) クラスタ内のタグは重心からの距離が最も近いタグ 4 種類および Intrusion tag はクラスタ外からランダムに 1 種類を選択する．

[*3] https://www.lancers.jp/



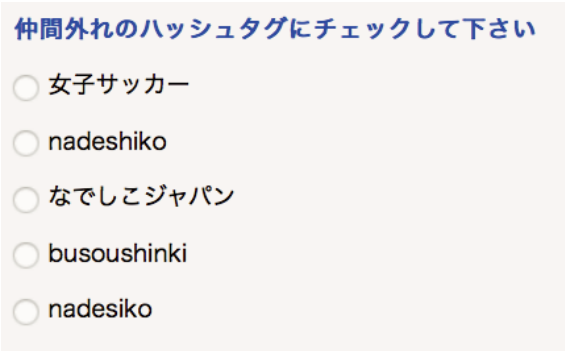


図 4　Lancers で行った Tag intrusion のタスクの一例

Fig. 4　An example of task of Tag intrusion by Lancers.

(2) 同時に表示するツイートは文字数が 28 から 51 であり，これを満たすものからランダムに 1 ツイート表示する．

(3) ハッシュタグのタイトルの長さは 23 文字以内．

(1) の条件から，評価に用いるクラスタは，含まれるタグが少なくとも 4 つ以上存在するクラスタとなる．(2) の条件において 28 は全ツイートの文字数の中央値，51 は第 1 四分位数である．(3) の条件における 23 文字は全ハッシュタグのタイトルの長さの累積相対度数が 0.95 のときの度数である．以上の条件を満たすクラスタからランダムに 200 クラスタをサンプリングし，それぞれのクラスタに対して 5 タスク作成し，一手法あたり計 1,000 タスクを用意した．また同クラスタにおいてタスクごとに表示するツイートはランダムに選択される．実際に行ったタスクの例を**図 4** に示す．それぞれのタスクではユーザに対して，まずハッシュタグとそのタグに対応するツイートを確認して，これらのツイートからユーザが仲間はずれだと考えるタグを Intrusion tag としてラジオボタンで選択してもらう．この例では，「#busoushinki」が Intrusion tag である．本研究ではこれらのタスクに対するユーザ検証の結果から，Intrusion tag の全体の選択回数の集計と 1 クラスタにおける Intrusion tag の選択回数の集計を行い，手法間のクラスタリング結果が人にとって分かりやすくまとまっているかを評価する．

4.6　実行時間の比較方法

本研究では，提案手法とベースラインにおいてそれぞれに独立した処理における実行時間の比較を行う．提案手法ではバーストタグの検出，バーストタグのクラスタリング，非バーストタグの割り当てを一連の処理としており，この時間の計測を行う．ベースラインでは，ハッシュタグのバースト検出，バーストタグと非バーストタグを組み合わせたハッシュタグ集合のクラスタリングを一連の処理としており，この時間の計測を行う．実行時間は，それぞれ処理の開始から終了までの時間の計測を 10 回行いその平均実行時間を比較に用いる．

5.　実験結果

5.1　Tag intrusion によるクラスタのまとまりのよさ

提案手法によるクラスタ数 10%におけるバーストタグのクラスタリング結果と割り当てられた非バーストタグについて，2012 年ロンドンオリンピックの体操競技に関するクラスタを**表 2** に示す．表 2 では，体操競技に関するバーストタグとして「体操」や「内村航平」といったタグが所属しており，当該期間に体操競技の話題が流行していたと推定できる．そしてフィルタリングされなかった非バーストタグでは「olympic_gim」や「床」といった体操競技に関連したタグが割り当てられる．一方で「にわか」や「笑っていいとも」など一見して関連が低そうなタグがフィルタリングされていることが結果から分かる．

以上のクラスタを含む提案手法によって獲得したハッシュタグのクラスタに対して，Tag intrusion を行った結果を**表 3**，**図 5** に示す．表 3 は全選択回数中 Intrusion tag の選択された割合を示したものである．また χ 二乗検定における p 値は，ユーザ検証の結果がランダムに選択した場合と同等であるという帰無仮説を検定したものであり，有意確率 0.10%において帰無仮説が棄却された．比較の結果，提案手法の方が高い割合で Intrusion tag が選択されている．この結果を受けて，Intrusion tag の選択される割合に有意水準 5%において有意差があるか検証を行った．検証方法は，各クラスタにおける Intrusion tag の平均選択率を手法ごとに計算し，その結果として得られる 200 個の評価値の集合から，手法ごとに 2 つとり出し，両者が同じ分布に従うという帰無仮説を検定する．この検証において算出された評価値集合のヒストグラムを図 5 に，統計量を**表 4** に示す．ここで Intrusion tag の選択率は 0 以上 1 以下の実数値であり，正規分布に従う保証がないことから，本研究の検証では Bruuner-Munzel 検定 [21] を用いる．この検定はデータの正規性と等分散性を仮定しないノンパラメトリック検定であり，2 群のデータは同じ分布に従うと

表 2　クラスタ数 10%において提案手法による 2012 年ロンドンオリンピックの体操競技に関するクラスタ

Table 2　The cluster on 2012 London Olympic Gymnastics Games (k: 10%).

バーストタグ	フィルタリング	非バーストタグ	重心とのコサイン類似度
体操		sari_gorin	0.527
男子体操	対象外	stv	0.401
ArtisticGymnastics		olymic_gim	0.371
体操個人総合		床	0.350
taisou		デイリースポーツ	0.304
内村航平		NEWS ポストセブン私のイメージ名字	0.195
gymnastics		なんですか	0.180
内村		にわか	0.166
金メダル		tanakareina	0.151
		ブラサン食べて	
さぽスポ_五輪	対象	内村選手を応援するよ	0.135
女子個人総合		田中れいな	0.131
メダル欲しい		7ji	0.118
etyping		soKKuri	0.105
et_t		笑っていいとも	0.083
		HFsms	0.078
		モー娘	0.064

表 3　全選択回数中 Tag intrusion による Intrusion tag の選択された割合

Table 3　The selected percentage of Intrusion tags among all selections by tag intrusion.

手法	クラスタ数（%）	選択されたIntrusion tag の割合	χ 二乗検定による p 値
提案手法	10	0.646	0.000
	30	0.671	0.000
ベースライン	10	0.604	0.000
	30	0.651	0.000

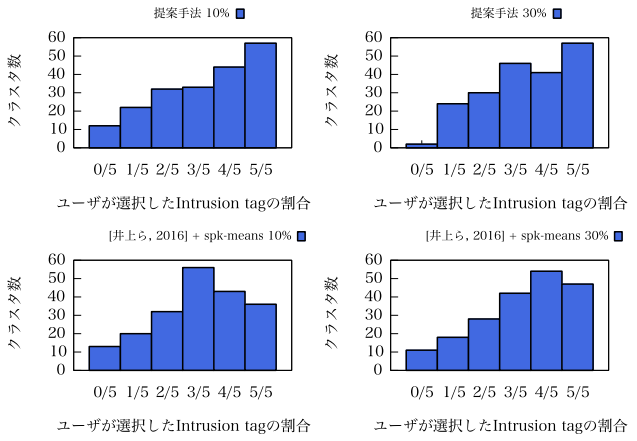


図 5　クラスタごとの Intrusion tag が選択される割合

Fig. 5　The selected proportions of intrusion tag for each cluster.

表 4　各手法におけるクラスタごとの Intrusion tag の選択される割合の集合の統計量

Table 4　The statistics of the set of selected proportions of intrusion tags for each cluster in each method.

手法	クラスタ数（%）	平均値	中央値	最頻値
提案手法	10	0.646	0.800	1.000
	30	0.671	0.600	1.000
ベースライン	10	0.604	0.600	0.600
	30	0.651	0.800	0.800

表 7　各手法におけるクラスタごとの誤選択された非バーストタグの割合の集合の統計量
Table 7　The statistic of the set of the proportion of misselected non-burst tags for each cluster.

手法	クラスタ数 (%)	バースト：非バースト								
		1:3			2:2			3:1		
		平均値	中央値	最頻値	平均値	中央値	最頻値	平均値	中央値	最頻値
提案手法	10	0.250	0.200	0.200	0.150	0.000	0.000	0.127	0.000	0.000
	30	0.257	0.300	0.400	0.152	0.100	0.000	0.074	0.000	0.000
ベースライン	10	0.269	0.200	0.200	0.191	0.200	0.200	0.148	0.000	0.000
	30	0.337	0.200	0.200	0.192	0.200	0.000	0.141	0.000	0.000

表 5　手法間の Bruuner-Munzel 検定結果
Table 5　The p-value of the Bruuner-Munzel test of each method.

提案手法	ベースライン	p 値
クラスタ数 10%	クラスタ数 10%	0.101
	クラスタ数 30%	0.958
クラスタ数 30%	クラスタ数 10%	0.025
	クラスタ数 30%	0.600

いう帰無仮説を検定する．この Bruuner-Munzel 検定の結果を**表 5** に示す．有意水準 5%で仮説が棄却されたのは提案手法のクラスタ数 30%かつベースラインのクラスタ数 10%だけであり各統計量が提案手法の方で高い値を示していることが分かったが，それ以外では有意差が確認されなかった．一方で，ベースラインに対して提案手法の評価値が優位に低いという結果が確認されなかったことから，その後の処理時間の計測結果で提案手法の処理時間の短縮が確認された場合，ベースラインと同等のタグのまとまりの良さがある状態で処理時間の短縮ができることを示唆している．

次にバーストタグのクラスタに対して非バーストタグが適切に割り当てられているか確認を行う．この検証では前述の検証と同様に Tag intrusion によって各クラスタで誤選択された非バーストタグの選択される割合を手法ごとに計算し，その結果として得られるクラスタの非バーストタグの選択の割合の集合から，手法ごとに 2 つとり出し，有意水準 5%において両者が同じ分布に従うという帰無仮説の検定を行った．この際，各クラスタのタスクに含まれる Intrusion tag ではないバーストタグあるいは非バーストタグの数はそれぞれ，バーストタグ：非バーストタグの比で (0 : 4)，(1 : 3)，(2 : 2)，(3 : 1)，(4 : 0) の場合が考えられる．仮にそれぞれの場合分けを考慮せず非バーストタグ選択される割合の計算を行った場合，タスクに含まれるバーストタグと非バーストタグの偏りに影響されて選択されたのか，あるいはクラスタのまとまりが悪いため選択されたのか評価することが困難である．よって本研究ではこれらの場合を分けて評価を行う．このうち (0 : 4)，(4 : 0) の割合でタグが偏るクラスタではバーストタグと非バーストタグの選択のされやすさに影響しないため，これ以外の三種類の場合にわけ，割合の集合を作り，それぞれを Bruuner-Munzel 検定によって提案手法とベースラインで割合の集合に有意差があるか検証を行った．

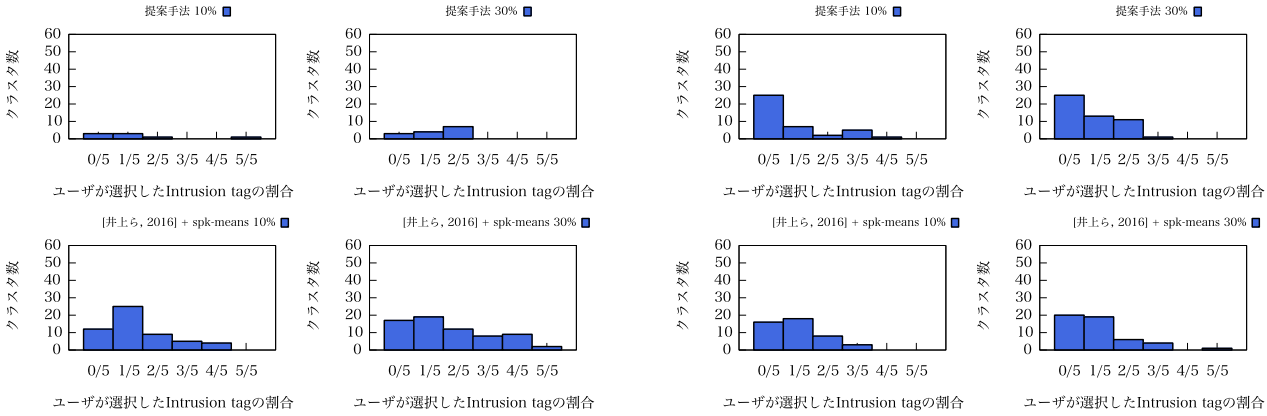
表 6　各手法かつバーストタグと非バーストタグの比ごとのクラスタの数
Table 6　The number of clusters in the ratio of each burst tag to non-burst tag.

手法	クラスタ数（%）	バースト：非バースト				
		0 : 4	1: 3	2 : 2	3 : 1	4 : 1
提案手法	10	2	8	40	66	84
	30	4	14	50	76	56
ベースライン	10	60	55	45	23	17
	30	39	67	50	34	10

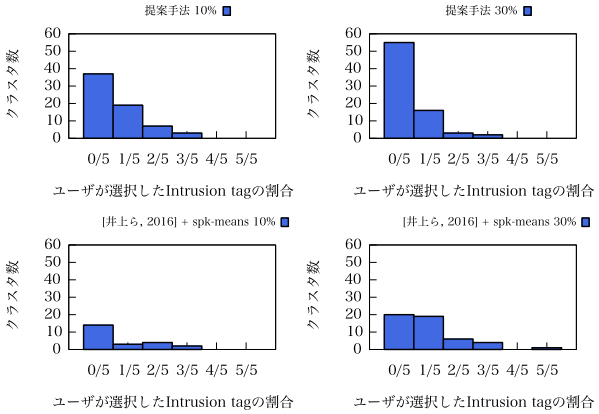
表 8　手法間のユーザが誤選択した非バーストタグの割合の集合の Brunner-Munzel 検定の p 値
Table 8　The p-value of the Brunner-Munzel test of the set of proportions of misselected non-burst tags.

提案手法	ベースライン	バースト：非バースト		
		1: 3	2 : 2	3 : 1
クラスタ数 10%	クラスタ数 10%	0.559	0.086	0.950
	クラスタ数 30%	0.378	0.132	0.753
クラスタ数 30%	クラスタ数 10%	0.737	0.273	0.212
	クラスタ数 30%	0.512	0.431	0.049

この検証におけるバーストタグと非バーストタグの比ごとに分けた非バーストタグの選択される割合の集合のヒストグラムを**図 6** の (a) から (c) に，各手法および各クラスタにおける非バーストタグの比ごとのクラスタ数を**表 6** に，それぞれの割合の集合の統計量を**表 7** に示す．そして，手法間の Bruuner-Munzel 検定の p 値を**表 8** に示す．検定結果において，有意水準 5%で帰無仮説が棄却されたのは非バーストタグにおいてタグ比 (3 : 1) の場合に提案手法のクラスタが 30%とベースラインのクラスタ 30%の組み合せであり，それ以外では各手法でそれぞれ同じ分布を持つ可能性があることが分かった．この組合せの統計量を比較すると，タグ比 (3 : 1) に場合に提案手法クラスタが 30%とベースラインがクラスタ 30%の組合せでは，中央値と最頻値が 0.000 で同一だが提案手法の方が平均値が低い



(a) 1 タスクにつきバーストタグ 1 つと非バーストタグ 3 つを含む (b) 1 タスクにつきバーストタグ 2 つと非バーストタグ 2 つを含む



(c) 1 タスクにつきバーストタグ 3 つと非バーストタグ 1 つを含む

図 6　バーストタグと非バーストタグがタスクに含まれる比ごとに場合分けを行ったクラスタについてユーザが誤選択した非バーストタグの割合の集計結果．(c) における提案手法 30%とベースライン 30%以外で評価値の有意差が確認されなかったことから，提案手法はベースラインと同程度の非バーストタグのまとまりの良さがあることが示唆される．

Fig. 6　The result of the proportion of non-burst tags missselected by the user. We divided case by burst tag and non-burst tag by ratio included in task. Since there is no significant difference in evaluation values except (c), we suggest that the proposed method has good cohesion of non-burst tags at the same level as the baseline.

値となっており，この組合せにおいては提案手法の方が非バーストタグが選択されにくいことが分かった．以上の結果から，Intrusion tag の場合と同様に，ベースラインに対して提案手法の評価値が優位に低いという結果が確認されなかったことから，ベースラインと同等以上のタグのまとまりの良さがある可能性があることが分かった．

次に非バーストタグの割当てをクラスタリングのまとまりの良さの観点で評価するため，各手法におけるクラスタ内分散の平均値を求めた．本研究ではクラスタ内分散を，1 からクラスタ重心と所属ハッシュタグのコサイン類似度を引いた値を各クラスタごとに計算を行い，それらの平均が小さいほどクラスタ重心にハッシュタグが集中していると評価する．**表 9** より，提案手法はベースラインと比較してクラスタ内分散の平均値が高く，クラスタ内のまとまりが悪いことが分かった．また非バーストタグの割り当て前後でクラスタ内分散の平均値が高くなっており，非バーストタグの割り当て処理がクラスタのまとまりの良さに悪影響を与えていることが分かった．以上の結果から，クラスタ重心から近いタグに対してユーザ検証を行う Tag intrusion では提案手法がベースラインと同等以上のまとまりの良さがある可能性が高い一方で，クラスタ全体を評価するクラスタ内分散の平均値はベースラインより低い値となっている．さらに非バーストタグの割当て前後でクラスタ内分散の平均値が低くなっていることから，クラスタ重心から離れた非バーストタグの割り当ては提案手法の方がまとまりが悪くなる可能性があることが分かった．

5.2　実行時間の比較結果

両手法における実行時間の結果を**表 10** に示す．提案手法による実行時間の合計は，クラスタ数 10%，30%において

表 9　各手法におけるクラスタ内分散の平均値
Table 9　The average value within-claster variance in each method.

手法	クラスタ数（%）	非バーストタグ割当て前のクラスタ内分散の平均値	クラスタ内分散の平均値
提案手法	10	0.205	0.649
	30	0.053	0.320
ベースライン	10	-	0.167
	30	-	0.040

表 10　各手法における平均実行時間［秒］
Table 10　The average execution time in each method [sec].

手法	クラスタ数 (%)	クラスタ数 (k)	ベクトルの次元数	バーストタグの検出	クラスタリング	非バーストタグの割当て	実行時間の合計
提案手法	10	603	32,414		114	46	9,312
	30	1,808	32,414	9,152	241	134	9,527
ベースライン	10	2,416	86,972		3,326	-	12,478
	30	7,248	86,972		9,058	-	18,120

それぞれを分換算した場合，156 分，159 分であった．一方でベースラインではそれぞれ 208 分，302 分必要であり，提案手法による実行時間の短縮が確認できた．また最も時間がかかった処理はバーストタグの検出であることが分かった．この点に関して，ハッシュタグのバースト検出は各タグごとに処理が独立していることから容易にスケーリングできるため，さらに実行時間を短縮できる可能性がある．そして全体の処理で比較を行った場合，提案手法がベースラインよりも実行時間が短くなっていることが分かった．

6. おわりに

本研究では，現実世界で起こった出来事や注目されている話題を Twitter 上の投稿から検出することを目的として，バーストタグとそれらのタグに関連性のある非バーストタグによって形成されたハッシュタグクラスタを生成する手法を提案した．ここでは，一定期間にバーストタグをクラスタリングし，非バーストタグをそのクラスタに割り当てる手法を提案した．実験により，Twitter ハッシュタグから Tag Intrusion によって提案手法の評価を行い，それぞれのクラスタごとにクラスタ重心に近いハッシュタグでは話題のまとまりの良さを保ちつつ，実行時間が短縮されることが分かった．

一方でベースラインと比較した際に，クラスタ全体におけるバーストタグと非バーストタグのまとまりは悪く，改善を行う必要がある．このまとまりの悪さの原因について，提案手法では各クラスタで同じベータ分布を用いた非常に粗い近似を用いたフィルタリングを行っており，クラスタごとのデータ点の集中度合いの考慮していないことにあると考えられる．この問題に対するアプローチとして，クラスタリング手法をデータ点の集中度を考慮した手法である混合 von Mises-Fisher 分布に変更することが考えられる．さらにこの手法ではデータ点が任意のクラスタに所属する確率を求めることができるため，ベータ分布によるハッシュタグとクラスタ重心の距離の分布の点推定が必要なくなり，クラスタリングと一貫したハッシュタグの割当てを行うことができる．今後は混合 von Mises-Fisher 分布によるクラスタリング結果に対して，同様のユーザ検証と実行時間の計測を行い，非バーストタグのまとまりの悪さの改善と実行時間の変化を確認する．

謝辞　本研究の一部は，JSPS 科研費（課題番号 16H02904）および筑波大学図書館情報メディア系プロジェクト研究の助成によって行われた．

参考文献

[1] Wallop, H.: Japan earthquake: How Twitter and Facebook helped - Telegraph (2011), available from ⟨https://www.telegraph.co.uk/technology/twitter/8379101/Japan-earthquake-how-Twitter-and-Facebook-helped.html⟩.
[2] 芥子育雄，鈴木　優，吉野幸一郎，大原一人，向井理朗，中村　哲：単語・パラグラフの分散表現を用いた Twitter からの日本語評判情報抽出，第 8 回データ工学と情報マネジメントに関するフォーラム (2016).
[3] 荒牧英治，増川佐知子，森田瑞樹：Twitter Catches the Flu：事実性判定を用いたインフルエンザ流行予測，研究報告音声言語情報処理（SLP），Vol.2011, No.1, pp.1–8 (2011).
[4] 水沼友宏：Twitter におけるバーストの生起要因と類型化に関する分析，修士論文，筑波大学 (2014).
[5] Diao, Q., Jiang, J., Zhu, F. and Lim, E.-P.: Finding Bursty Topics from Microblogs, *Proc. 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, *ACL '12*, pp.536–544, Association for Computational Linguistics (2012).
[6] Du, Y., Wu, W., He, Y. and Liu, N.: Microblog bursty feature detection based on dynamics model, *2012 International Conference on Systems and Informatics* (*ICSAI2012*), pp.2304–2308 (2012).
[7] Guozhong, D., Ruiguang, L., Wu, Y., Wei, W., Liangyi, G., Guowei, S., Miao, Y. and Jiguang, L.: Microblog Burst Keywords Detection Based on Social Trust and Dynamics Model, *Chinese Journal of Electronics*, Vol.23, No.4 (2014).
[8] 木村　輔，宮森　恒：共起と潜在トピックを考慮したハッ

シュタグ間関係の分類手法，電子情報通信学会論文誌，Vol.J98-D, No.8, pp.1151–1161 (2015).
[9] Tsur, O., Littman, A. and Rappoport, A.: Efficient clustering of short messages into general domains, pp.621–630 (2013).
[10] 井上優作，若林　啓：表記の多様性を考慮したハッシュタグ推薦，第 14 回日本データベース学会年次大会 (2016).
[11] 福山怜史，若林　啓：バースト現象を考慮したハッシュタグのクラスタリング手法の提案，研究報告情報基礎とアクセス技術（IFAT），Vol.2017-IFAT-128, No.17, pp.1–6 (2017).
[12] Kadota, K., Ye, J., Nakai, Y., Terada, T. and Shimizu, K.: ROKU: An Improved Method for the Detection of Tissue-Specific Expression Patterns (2006).
[13] 石川栄介：棄却検定の比較表，岩手大学学芸学部研究年報，Vol.15, No.2, pp.1–7 (1960).
[14] Sprent Peter, S.N.A.: Applied Nonparametric Statistical Methods, *Chapman and Hall*, p.480 (1993).
[15] Peter, S.: Data Driven Statistical Methods, *Chapman and Hall*, p.406 (1997).
[16] Dhillon, I.S. and Modha, D.S.: Concept Decompositions for Large Sparse Text Data Using Clustering, *Machine Learning*, Vol.42, No.1, pp.143–175 (2001).
[17] 井出　剛，杉山　将：異常検知と変化検知，講談社 (2015).
[18] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L. and Blei, D.M.: Reading Tea Leaves: How Humans Interpret Topic Models, *Advances in Neural Information Processing Systems 22*, Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I. and Culotta, A. (Eds.), pp.288–296, Curran Associates, Inc. (2009).
[19] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, *Proc. EMNLP*, pp.230–237 (2004).
[20] Arthur, D. and Vassilvitskii, S.: K-means++: The Advantages of Careful Seeding, *Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, *SODA '07*, pp.1027–1035, Society for Industrial and Applied Mathematics (2007).
[21] Edgar, B. and Ullrich, M.: The Nonparametric Behrens: Fisher Problem: Asymptotic Theory and a Small-Sample Approximation, *Biometrical Journal*, Vol.42, No.1, pp.17–25 (2000).

福山 怜史

2017 年福井大学工学部情報・メディア工学科卒業．現在，筑波大学大学院図書館情報メディア研究科博士前期課程在学中．ソーシャルメディアに焦点を当てたデータマイニングの研究に従事．

若林 啓

2012 年法政大学大学院工学研究科博士課程修了．博士（工学）．同年筑波大学図書館情報メディア系助教．機械学習の研究に従事．電子情報通信学会，日本データベース学会，ACM 各正会員．

（担当編集委員　是津 耕司）