Vol.2017-DBS-165 No.17
Vol.2017-IFAT-128 No.17
2017/9/19

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

福山 怜史[1,a)]　若林 啓[2,b)]

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上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抽象度が曖昧な性質が存在するため，同じ話題を指し示しているタグが重複したり，バーストしていないタグでもバーストしたタグと同じ話題を示すタグが存在する可能性がある．この問題を解決するためにハッシュタグのクラスタリングが必要であるが，全てのタグのクラスタリング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，バーストしたハッシュタグのみクラスタリングを行い，得られたクラスタにバーストしていないハッシュ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のサンプル数はバーストしたハッシュタグだけになるため，クラスタリングに掛かる処理時間が短縮できる．提案手法では，移動平均線収束拡散法によってハッシュタグのバーストの強さを数値化し，異常値を持つとして検出されたハッシュタグをバーストしたハッシュタグとみなす．そして，バーストしたハッシュタグを k 平均法によってクラスタリングし，バーストしていないハッシュタグを生成したクラスタに割り当てる．本稿では，先行研究の手法によって生成されたクラスタと比較して，提案手法によるクラスタが，話題に対するハッシュタグの網羅性が高く，クラスタリングにかかる処理時間が大幅に短縮されることを報告する．

キーワード：Twitter，ハッシュタグ，バースト，移動平均線収束拡散法，クラスタリング

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．例えば，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]．また，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，Twitter のリアルタイム性を利用した分析の前処理などにおいて有用である．

Twitter では，ユーザは特定の話題についての言及であ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“バーストしたハッシュタグ”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，単純にバーストしたハッシュ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，バーストしたハッシュタグの集合の中には，同じ話題を指し示

[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
[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
[a)] s1721691@s.tsukuba.ac.jp
[b)] kwakaba@slis.tsukuba.ac.jp

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，バーストしていないハッシュタグの中にも，バーストしたハッシュタグと同じ話題を指し示している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．例えば，体操競技に関係する「#体操」というハッシュタグがバーストしている時に，バーストしていないが関連したハッシュタグ（特定の選手に注目した「#内村航平」など）を抽出することは，当該の話題の全容を把握する上で重要である．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．井上ら [9]，木村ら [10] は，共起する単語を特徴量としてハッシュタグのクラスタリングを行うことによって，表記揺れや構造関係を吸収し，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全てのハッシュタグのクラスタリングを行った上で，バーストしたハッシュタグを含むクラスタを出力する必要がある．しかし，クラスタリング処理ではクラスタ数とサンプル数の積に比例して処理時間が増加するため，全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，バーストしたハッシュ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．

本研究では，バーストしたハッシュタグのみを用いてクラスタリングを行い，バーストしていないハッシュ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したハッシュタグのみを考慮すれば良いため，クラスタリングにかかる処理時間を大幅に短縮できる．また，全てのハッシュタグをクラスタリングする場合に比べて，バーストした話題に対してよりまとまりの良いクラスタが生成されることも期待できる．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全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，抽出されるハッシュタグの網羅性が向上し，処理時間が大幅に短縮されることを示す．

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や構造化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] と Du ら [6]，Guozhong ら [7] の研究がある．水沼ら [4] 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なっている．Du ら [6]，Guozhong ら [7] は，Twitter においてバーストした話題を抽出するために，移動指数平均線平滑法を用いた手法を提案している．Du ら [6] は，Twitter にてバーストした語，バーストキーワードを検出するためにリツイート数や投稿日時を考慮し，ツイートに出現する語のバーストの強さを計算する手法を提案している．Guozhong ら [7] は，Du ら [6] の手法においてバーストの誤検出を防ぐために，ユーザの信頼性と，過去数日の同時刻帯の語の出現頻度を考慮した手法を提案している．本研究では，バーストしたハッシュタグの検出に，Du ら [6]，Guozhong ら [7] の用いた移動指数平均線平滑法によってハッシュタグのバーストの強さを計算し，その値からバーストの検出を行う．

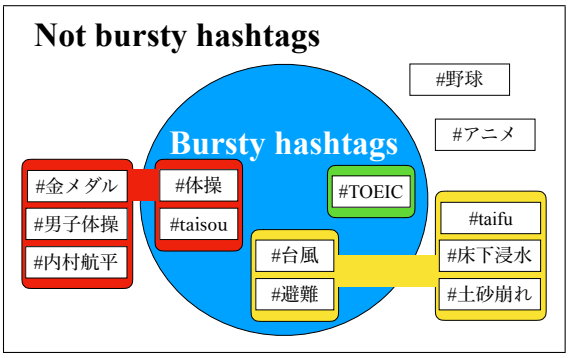

図 1　ハッシュタグと Twitter で流行している話題の関係

次にハッシュタグのクラスタリングや構造化に関する研究として，木村ら [10]，丹羽ら [8] によるハッシュタグの構造化に関する研究，Tsur ら [11]，井上ら [9] によるハッシュタグクラスタリングに関する研究がある．ハッシュタグの構造化に関する研究では，ハッシュタグの関係をあらかじめ仮定したいくつかのクラス（同義，階層，関連，関連なし）に分類する手法が提案されている．丹羽ら [8] の研究は，ハッシュタグのようにユーザが自由にタグを作成することのできるフォクソノミーの分析と構造化手法の提案を行う研究である．ハッシュタグクラスタリングの研究では，Tsur ら [11] によって事象ごとにハッシュタグをクラスタリングする手法が提案されている．ここでは，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全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 [11] 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 平均法を用いている．また，井上ら [9] は，Tsur ら [11] が英語で行な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を示した．本研究では，Tsur ら [11]，井上ら [9] と同様に，ハッシュタグ毎にツイートを全て結合した擬似文書から TF-IDF ベクトルを作成し，k 平均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，バーストしたハッシュタグの検出，バーストしたハッシュタグのクラスタリング，生成したクラスタからバーストしていないハッシュタグの割り当てを行う．

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移動平均線収束拡散法 (moving average convergence divergence, MACD) を用いる．移動平均線収束拡散法とは，トレンド分析手法の 1 つで，短期 s と長期 l の 2 つの異なる期間を設定し，その期間の移動指数平均の差異から対象となるサイズ $n+1$ の時系列データ $\boldsymbol{x} = (x_0, x_1, ..., x_t, ..., x_n)$ のトレンドを定量化する手法 [12] である．移動平均線収束拡散法による定量化したトレンドの強さの導出を述べる．まず短期と長期における 2 種類の移動指数平均 (EMA) を求める．時刻 t における $n \leq t$ を満たす期間 n の場合の移動指数平均を算出する式を (1) に示す．

$$EMA_t^{(n)} = \sum_{k=0}^{n} \alpha_n (1-\alpha_n)^k x_{t-k} \tag{1}$$

$$\alpha_n = \frac{2}{n+1} \tag{2}$$

(2) は，平滑化定数である．時刻 t における短期間を s，長期間を l とした時の移動指数平均から，式 (3) を算出する．

$$MACD_t^{(s,l)} = EMA_t^{(s)} - EMA_t^{(l)} \tag{3}$$

次にこの $MACD_t^{(s,l)}$ の変動量である $MACD$ ヒストグラムを計算する．$MACD$ ヒストグラムは，$s < m < l \leq t$ を満たす m をパラメータとした移動指数平均より，$MACD_t^{(s,l)}$ の値を式 (4) により平滑化し，元の $MACD_t^{(s,l)}$ との差分として式 (5) によって計算される．

$$signal_t^{(s,m,l)} = \sum_{k=0}^{m} \alpha_m (1-\alpha_m)^k MACD_{t-k}^{(s,l)} \tag{4}$$

$$hist_t^{(s,m,l)} = MACD_t^{(s,l)} - signal_t^{(s,m,l)} \tag{5}$$

本研究ではこの $MACD$ ヒストグラムを用いて，ハッシュタグのバーストの強さを次の手順で定量化する．はじめに Twitter における任意のハッシュタグ h が付与されたツイートの単位時間あたりの出現頻度を f とする．これにより，ハッシュタグ H の出現頻度のデータセットは時系列順に $F_h = (f_{h,0}, f_{h,1}, ..., f_{h,t}, ..., f_{h,n})$ となる．このデータセットから時系列順にデータを取り出し，$MACD$ ヒストグラム (5) を計算する．本稿では，ハッシュタグ h における $MACD$ ヒストグラムによるデータセットを BST_h，ハッシュタグ h における $hist_t^{(s,m,l)}$ を $BST_{h,t}$ とし，$BST_h = (BST_{h,0}, BST_{h,1}, ..., BST_{h,t}, ..., BST_{h,n})$ と表記する．

次にこの求められたデータセットから，標準偏差を比較する方法によって，バーストの発生を検出する．この手法ではスライディングウィンドウ方式を採用し，$W_{h,t-1}$，$W_{h,t}$ の 2 つデータセットを比較し，バーストの発生を検出する．

$$W_{h,t-1} = (BST_{h,t-n'}, BST_{h,t-n'+1}, ..., BST_{h,t-1})$$
$$W_{h,t} = (BST_{h,t-n'+1}, BST_{h,t-n'+2}, ..., BST_{h,t})$$

n' をウィンドウのサイズとする．$W_{h,t-1}$ は新たなデータ $BST_{h,t}$ が到達する前のデータセット，$W_{h,t}$ は $BST_{h,t}$ が到着した時のデータセットである．スライディングウィンドウ方式より $BST_{h,t}$ が到達後のデータセット $W_{h,t}$ では $W_{h,t-1}$ の先頭のデータ $BST_{h,t-n}$ が取り除かれ，末尾に新しいデータ $BST_{h,t}$ が挿入されている．以上のデータセットを用いて，データセット $W_{h,t}$ が $W_{h,t-1}$ の標準偏差より 3 倍以上大きくなっている場合，時系列データの挙動が変化し異常値として高い値を示しているため，バーストが発生しているとみなす．

3.2 ハッシュタグのクラスタリング

バーストしたハッシュタグの集合を 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．

$$tf(w_i, h) = \sum_{d_h \in D_h} tf(w_i, d_h) \tag{6}$$

また，語彙 w_i の文書頻度を以下のように定義する．

$$df(w_i) = \sum_{h \in H'} U(tf(w_i, h) - 1) \tag{7}$$

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．

$$idf(w_i) = \log \frac{1 + |H'|}{1 + df(w_i)} + 1 \tag{8}$$

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 $\boldsymbol{v}'_h = (v_{h,1}, ..., v_{h,|W|})$ 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．

$$v_{h,i} = tf(w_i, h) \cdot idf(w_i) \tag{9}$$

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．

$$\boldsymbol{v}_h = \frac{\boldsymbol{v}'_h}{\|\boldsymbol{v}'_h\|} \tag{10}$$

以上の方法で全てのハッシュタグにおけるベクトルを作成し，k 平均法を用いてクラスタリングを行う．k 平均法への入力のデータセットは $\boldsymbol{v}_h$ を使用し，対応するハッシュタグのクラスタを推定する．

3.3 バーストしていないハッシュタグのクラスタへの割り当て

3.2 節で生成したクラスタに対して，バーストしていないハッシュタグを割り当てる方法について述べる．まず，バーストしていないハッシュタグの 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，文書頻度の値は変化していないと仮定し，特徴量を抽出する語彙と IDF の値はバーストしたハッシュタグの特徴量と同じものに設定する．次に，作成されたバーストしていない各ハッシュタグの特徴量ベクトルから，各クラスタが持つセントロイドの位置までのユークリッド距離を計算し，最近傍のクラスタを選択する．そして，この選択されたクラスタを特徴量ベクトルが示すハッシュタグのクラスタとみなす．

4. 評価実験

本研究の提案する手法を評価するために，Twitter のツイートに対して，以下の点について確認を行った．

(1) バーストしたハッシュタグにおいて重複した内容を持つタグの組み合わせの存在の確認
(2) バーストしていないハッシュタグの割り当て結果の妥当性の確認
(3) 先行研究である井上ら [9] の手法と比較した，処理時間短縮の確認

(1) では，バーストが検出されたハッシュタグ集合において，重複した話題を持つタグが存在するか確認を行う．もし重複した内容のタグが存在しない場合，提案手法においてバーストが検出されたハッシュタグ集合のクラスタリングをする必要性が失われる．よって，重複した内容のタグが存在するか検証を行い，クラスタリングの必要性を検討する．(2) では，提案手法で生成したクラスタに対して，バーストしていないハッシュタグのクラスタを割り当てた場合，この割り当てが問題なく行われるか検証を行う．(3) では，先行研究である井上ら [9] の手法を素直に適用して，全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，処理時間がどの程度短縮されたかについて確認する．この際，処理時間のボトルネックとなる k 平均法によるクラスタリングに掛かる時間を計測する．

4.1 実験データ及び環境

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のうち，その期間内で 100 件以上出現したタグ 18,377 件を対象とする．各ハッシュタグのコーパスに用いるツイートはハッシュタグと同様の期間に発生した 25,897,072 件のツイートを用いた．またバースト検出では，2012 年 7 月 17 日から 2012 年 8 月 7 日の期間における各ハッシュタグの出現頻度のデータを使用した．実験環境は，OS が Ubuntu 16.04，CPU が Intel Xeon E5-2630 (2.40GHz)8core 2 機，開発環境は Python で行った．

4.2 バースト現象の検出における設定

2012 年 8 月 1 日から 8 月 7 日の間に出現したハッシュタグ 18,377 件から，バースト現象が発生するハッシュタグの抽出を行った．バースト現象の検出方法は 3.1 節で述べた方法で，Du ら [6]，Guozhong ら [7] と同様のパラメータである $(s, m, l) = (4, 5, 8)$ を用いた．しかし日毎の出現頻度が極めて少ないハッシュタグは，偶然出現した場合にバースト現象が誤検出されてしまう恐れがある．そこで本研究では，バースト現象が検出された際に少なくとも 10 ツイート以上の出現頻度があることを条件とした．またバーストの周期性を考慮して，時系列の単位を 1 日，スライディングウィンドウのサイズを 7 に設定した．以上の条件でバースト検出を行なった結果，18,377 件のうち 1,933 件がバースト現象を経験したハッシュタグとして検出された．

4.3 ハッシュタグの特徴量ベクトルとクラスタリングの設定

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，k 平均法によってクラスタリングを行う．

提案手法におけるクラスタ数は，井上ら [9] の報告した最適なクラスタ数の 1 つである全体の 70%の 1,353 件とした．また先行研究である井上ら [9] の手法による比較対象データは，提案手法と同様の語彙の制約の上で，18,377 件のハッシュタグ全てを一括でクラスタリングを行う．この時のクラスタ数は，提案手法のクラスタ数と等しい 1,353

件，ハッシュタグ全体の 70%である 12,864 件の 2 種類でクラスタリングを行った．

本研究では，k 平均法の最初のセントロイドを kmeans++ 法 [13] によって，確率的に選択する．形態素解析器は MeCab[14]，TF-IDF および k 平均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[*1]を用いた．

5. 実験結果

5.1 クラスタリングとクラスタの割り当ての結果

提案手法による，バーストしたハッシュタグのクラスタリングとバーストしていないハッシュタグのクラスタへの割り当ての結果として，「体操」と「プリキュア」の 2 種類のハッシュタグを含むクラスタの例を表 1，表 2 に示す．表 1 では，バーストしたハッシュタグのクラスタリング結果から，体操競技に関する話題のクラスタが得られたと考えられる．そして体操競技の話題に関連するオリンピックや内村航平選手の話題のハッシュタグがこのクラスタに割り当てられたと考えられる．表 2 では，バーストしたハッシュタグのクラスタリング結果から，スマイルプリキュアというアニメ番組に関する話題のクラスタが得られ，その表記揺れに該当するハッシュタグやニチアサキッズタイムに関連するハッシュタグがこのクラスタに割り当てられたと考えられる．以上の例では，バーストしたハッシュタグにおいて重複した話題が存在することと，バーストしていないハッシュタグの割り当てが問題なく行われたことが確認された．

表 1 ハッシュタグ「体操」を含むクラスタ

クラスタリング結果 (バーストしたタグ)	クラスタへの割り当て結果 (バーストしていないタグ)
体操	ArtisticGymnastics
gymnastics	olimpic
taisou	Olympic2012
内村	yjfc_kohei_uchimura
	内村航平
	金メダル
	男子体操

表 2 ハッシュタグ「プリキュア」を含むクラスタ

クラスタリング結果 (バーストしたタグ)	クラスタへの割り当て結果 (バーストしていないタグ)
プリキュア	PreCure
precure	nichiasa
smile_precure	purecure
smile	smileprecure
スマイルプリキュア	スイプリ
	スマプリ
	ニチアサ

[*1] http://scikit-learn.org/stable/about.html#citing-scikit-learn

5.2 先行研究との比較

提案手法と先行研究による手法によって得られたクラスタの結果の比較を図 2 から図 5 に示す．ここでは，表 1，表 2 によって得られたハッシュタグが所属するクラスタを例として検証している．

図 2，図 3 より，提案手法によるクラスタと先行研究 ($k = 1,353$) によるクラスタで共通したハッシュタグが存在することがわかる．また共通していないハッシュタグに関しても，そのクラスタ内の話題に関係するハッシュタグであることから，提案手法と先行研究でクラスタが示す話題は同じであるといえる．

図 4，図 5 では，提案手法のクラスタは，先行研究 ($k = 12,864$) のクラスタを複数合併させた構造となっており，話題に対してクラスタの断片化がより少ないことを示している．このことから，提案手法の方がより冗長性が少なく，ハッシュタグ抽出の網羅性の高い出力を行うことができるといえる．

また，k 平均法の処理時間の比較結果を表 3 に示す．提案手法による処理時間は 10 分程度と最も短いのに対して，先行研究の手法では 18 時間以上の処理時間が必要であり，提案手法による大幅な処理時間の短縮が確認できる．

以上の結果から，バーストしたハッシュタグにおいて重複した内容を持つタグの組み合わせは存在し，バーストしていないハッシュタグの割り当ては一部の例において成功していることが確認できた．また，k 平均法の処理時間は先行研究の手法と比較して 100 倍以上短縮されていることがわかった．

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，バーストしたハッシュタグと関連するハッシュタグクラスタを抽出する手法を提案した．ここでは，一定期間にバーストしたハッシュタグをクラスタリングし，バーストしていないハッシュ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，実際の Twitter ハッシュタグから本手法の評価を行い，クラスタリングにかかる処理時間が 100 倍以上短縮され，一部の例において有用性の高いクラスタが得られたことを確認した．

今後の課題として，クラスタリングの有効性を一部のクラスタの主観的評価のみで行っている点が挙げられる．今後は，全てのクラスタに対して，バーストしていないハッシュタグの割り当てが適切に行われているかについて，クラウドソーシングなどを用いた統計的な検証を行う必要があると考えられる．

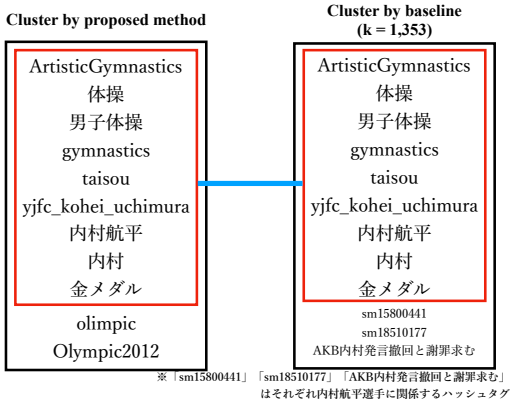

図 2 提案手法のクラスタ (表 1) と先行研究 ($k = 1,353$) のクラスタのハッシュタグが複数共通している例

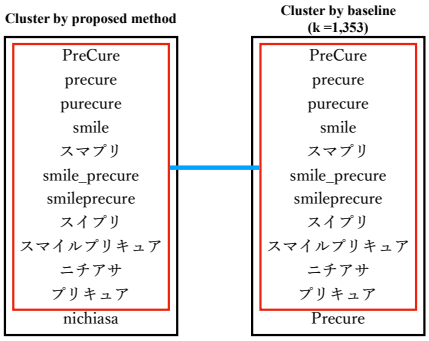

図 3 提案手法のクラスタ (表 2) と先行研究 ($k = 1,353$) のクラスタのハッシュタグが複数共通している例

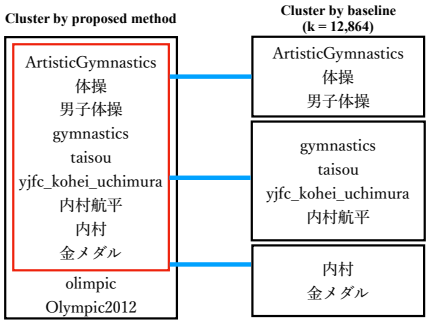

図 4 提案手法のクラスタ (表 1) が先行研究 ($k = 12,864$) のクラスタの複数合併となる例

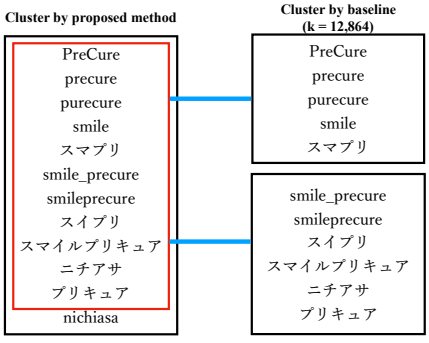

図 5 提案手法のクラスタ (表 2) が先行研究 ($k = 12,864$) のクラスタの複数合併となる例

表 3 各手法における処理時間

手法	クラスタ数	処理時間 [sec]
提案手法	1,353	611
先行研究	1,353	68,225
	12,864	224,975

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, *The Telegraph*, (online), available from ⟨http://www.telegraph.co.uk/technology/twitter/8379101/Japan-earthquake-how-Twitter-and-Facebook-helped.html⟩ (2011).

[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：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，Vol. 2011-SLP-86, pp. 1–8 (2011).

[4] 水沼友宏，　池内淳，山本修平，山口裕太郎，佐藤哲司，島田諭：Twitter におけるバーストの生起要因と類型化に関する分析，情報社会学会誌，Vol. 7, No. 2, pp. 41–50 (2012).

[5] Diao, Qiming, Jiang, Jing, Zhu, Feida and Lim, E.-P.: Finding Bursty Topics from Microblogs, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 536–544 (2012).

[6] Du, Y., Wu, W., He, Y. and Liu, N.: Microblog bursty feature detection based on dynamics model, *International Conference on Systems and Informatics*, 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] 丹羽智史，土肥拓生，本位田真一：Folksonomy の 3 部グラフ構造を利用したタグクラスタリング，*SIG-SWO-A602*, pp. 0701 – 0708 (2006).

[9] 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).

[10] 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).

[11] Tsur, O., Littman, A. and Rappoport, A.: Efficient Clustering of Short Messages into General Domains, *International Conference on Weblogs and Social Media (ICWSM)* (2013).

[12] ダヌシカ・ボレガラ，岡崎直観，前原貴憲：ウェブデータの機械学習，株式会社講談社 (2016).

[13] Arthur, D., Vassilvitskii and S: k-means++: The Advantages of Careful Seeding, *Society for Industrial and Applied Mathematics Philadelphia*, pp. 1027–1035 (2007).

[14] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis,, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237 (2004).