表記の多様性を考慮したハッシュタグ推薦

井上　優作†　　若林　　啓††

† 筑波大学情報学群知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1–2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †s1211468@u.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　本研究の目的は，現実世界で発生しているイベントの特徴をハッシュタグの仕組みを用いて学習し，SNS 上でそのイベントについて言及しているユーザをハッシュタグの推薦を通して特定することである．ハッシュタグを含むテキストを全て結合したものをそのハッシュタグの特徴を表す文書とし，そのような文書から成るコーパスについて各文書の TF-IDF を求め，ハッシュタグをクラスタリングする．そのクラスタとイベント時間中にユーザが投稿したテキストから求めた TF-IDF ベクトルとの類似度を計算してハッシュタグクラスタを推薦する．実験では，クラスタの中心よりも k-近傍法で推薦クラスタを決めたほうが精度が高かったことを示す．

キーワード　SNS, ソーシャル・ネットワーキング・サービス, ハッシュタグ, タグ付け

1. はじめに

Twitter(注1) や Facebook(注2) などのソーシャル・ネットワーキング・サービス (SNS) には日々大量にテキストが投稿されており，それらを系統的に整理して検索可能にすることは重要な課題である．特に，SNS に投稿される情報はリアルタイム性が強く，またユーザにしか知り得ない感想や情報が含まれることから，ライブや展示会，交通障害，災害，テレビ番組などといったイベントに関する情報源としての利用に注目が集まっている．

SNS から特定のイベントに関連した投稿を収集することができれば，当該イベントの主催者や当該イベントに関心のあるユーザにとって有益である．イベントと投稿を関連づける手段としては，関連するキーワードを用いる手法も考えられる [1] が，より直接的な方法として，ユーザが明示的にイベントとの関連を明らかにするために付与されるタグを用いることが考えられる．多くの SNS では，投稿の内容のテーマを表す文字列をタグとして付与することで，同様の内容の投稿を検索しやすくする機能がある．

例えば，Twitter においてこのような機能を果たす「ハッシュタグ」という仕組みは，ハッシュ記号 # の後に投稿の内容のテーマを示す文字列を続けることで,「#世界陸上」や「#人身事故」といったようにイベントに関連した投稿や実況であることの表明に利用されている．タグはこの他にも様々な用途で用いられる場合があるが，特に本稿ではイベントに関連したタグに着目し，イベントに関する情報収集に利用することを考える．

しかし，タグはユーザが自主的に入力して付与するものであるため，表記揺れやスペルミスなどによって，同じイベントについての投稿に対して異なるタグが使用されることが起こる．また，内容的に適切なタグが存在しているにも関わらず，ユーザがタグを付与していない投稿も多く存在していると考えられることから，タグによる検索は網羅性の観点から十分ではない．このため，タグに基づいてより網羅的にイベントに関連した投稿を収集するためには，同じイベントを指すと考えられるタグを整理し，さらにタグの付与されていない投稿に対してタグを推定するような手法が必要になる．

本研究では，特にあるイベントについてのタグについてその特徴を抽出し，ユーザの一定時間内の投稿群に対して付与できそうなタグを推定することで，そのユーザがイベントに参加しているかどうかを判定する手法を提案する．

具体的には，各タグごとにそのタグを含む投稿のテキストを全て結合した仮想文書を仮定し，そのような文書から成るコーパスを考える．このコーパスについて，各仮想文書の TF-IDF ベクトルをタグの特徴とする．つまり，本研究では，タグの付与されている投稿集合に現れる単語分布を特徴量としてタグのクラスタリングを行うことで，同じイベントを表すタグ集合を同定し，タグの付与されていない投稿についてイベントとの関連を推定する．

イベントに関連したタグには，同じタグでも実質的には異なる時間帯のイベントに用いられる特徴がある．例えば，毎週放送されるテレビ番組の実況を行うために用いられるタグは，毎週同じタグが使われるものの，各週で同じ内容ではないことから，そのタグの付与された投稿に現れるテキスト特徴は変化していくと考えられる．表 1 は，Twitter から 4 月 7 日，14 日，21 日の放送日にそれぞれ収集した「#precure」というテレビ番組に関するハッシュタグが付与された投稿に現れる単語ベクトルについて，4 月 7 日のハッシュタグとの類似度をそれぞれ求めた結果である．同じ放送日では同じ番組のハッシュタグ同士の類似度は高いが，翌週，翌々週の放送になるに従って，4 月 7 日の同じ番組のハッシュタグとの類似度は小さくなっていく．このことから，一週間などの一定の期間内に含まれる投稿のみをハッシュタグの特徴量とし，ユーザの投稿と同じ時間帯のハッシュタグとの類似度を求めることによって，ユーザの投稿とイベントとの関連を適切に推定できると考えられる．

(注1)：https://twitter.com
(注2)：http://www.facebook.com

4 月 7 日		4 月 14 日		4 月 21 日	
タグ	確率	タグ	確率	タグ	確率
precure	0.942	precure	0.605	precure	0.581
ドキプリ	0.813	nitiasa	0.483	nitiasa	0.431
ドキドキプリキュア	0.754	ドキプリ	0.460	ドキプリ	0.418
nitiasa	0.722	プリキュア	0.436	プリキュア	0.393
プリキュア	0.684	ドキドキプリキュア	0.410	ドキドキプリキュア	0.348
nichiasa	0.461	nichiasa	0.322	nichiasa	0.290
minako	0.075	黙れ	0.096	PrecureAll	0.079
SHT	0.062	寝言	0.091	寝言	0.077
ラフラフ	0.059	agqr	0.085	黙れ	0.076
tvasahi	0.055	違う	0.085	誰	0.074

表 1 『#precure』を含むツイートの各日付についてのハッシュタグ類似度

実際には，表記の上で互いに異なるタグが同じイベントを指している場合があるため，まずタグをクラスタリングし，投稿群に対してタグクラスタを推薦するのが望ましいと考えられる．本研究では，クラスタリングには k-means 法を使用する．その上で，タグクラスタの推薦手法として，k-means 法により得られた各クラスタの平均ベクトルまでの距離を用いる方法と，k 近傍法により得られた最も近いいくつかの仮想文書が所属するクラスタ番号の多数決を用いる方法を提案し，これらを実験により比較する．

2. 関 連 研 究

本研究に関わる先行研究で，領域が最も近いものは 2015 年の伊川らの研究 [1]，手法が最も近いものは 2013 年の Tsur らの研究 [2] である．

伊川ら [1] は，Twitter に投稿された特定のイベントについてのツイートを収集する時に，そのイベントについて多くの投稿をしたユーザのテキストからそのイベントに関わるキーワードと同時に出現する単語の特徴を学習することで，キーワードと同時に出現しかつイベントには関係無いキーワードをノイズキーワードと定義した．それを手がかりに，キーワードが出現するツイートの中からイベントに関係無いノイズツイートを高い精度で発見できることを示した．この研究ではイベントに関するキーワードは人手で定められていたが，本研究はイベントについてのタグを含むテキストから特徴を抽出することで，人手でキーワードを定義する必要が無くなる点に新規性がある．

Tsur ら [2] は，2011 年に Romero ら [3] が提案した Twitter のツイート分類方法に従うようにハッシュタグをクラスタリングすることを目指した．ハッシュタグごとにそのハッシュタグを含むツイートを全て結合した仮想文書を用意し，各文書から作成した特徴ベクトルを用いて文書をクラスタリングすることで，実質的にハッシュタグをクラスタリングすることになる．Tsur らは特徴ベクトルの作り方に TF-IDF ベクトル，またはハッシュタグの共起ベクトルを作り，クラスタリング手法には k-means 法 [4] を用いていた．この論文中では，ハッシュタグを k-means 法でクラスタリングする際にクラスタ数を 1,000 などの大きい数にすることで，似た意味を持つハッシュタグから成るクラスタや,「もし○○が××だったら」といった大喜利のお題とも言えるようなハッシュタグから成るクラスタが得られたことが示されている．本研究では，日本語のツイートを用いても同様の結果が高い精度で得られることを仮定し，得られたクラスタをさらに推薦の問題にも用いる．

2011 年に Antenucci らが行った研究 [5] では，ハッシュタグ同士の共起をグラフ構造とみなし，対象の 2 つのハッシュタグが互いにどれだけ強く共起しているかを類似度として様々な手法でクラスタリングを行い，その結果に対してツイートの分類をした．その中では，主成分分析によって文書の次元を削減したほうが分類の精度が高くなることが示されている．

2013 年の Godin らの論文 [6] は，Latent Dirichlet Allocation [7] を用いて Twitter のツイートに対してハッシュタグを推薦する手法を提案している．具体的には，1 ツイートを 1 文書と見なしてトピック分布を推論し，その分布から一定回数トピックをサンプリングし，サンプルされたトピックの中で，そのトピックに所属する確率が高い単語から順にハッシュタグとして推薦している．この手法は特にハッシュタグについての特徴を学習しているわけではないため，普通の単語をハッシュタグとして使用するよう推薦している．そのため，time, love, car といった漠然とした単語が多く推薦されるが，これらのような単語は意味が広すぎて，特定の話題を示すためのハッシュタグとしては使いづらいと考えられる．

2014 年の木村らの論文 [8] は，ハッシュタグ-ツイート本文中の単語-ユーザの三部グラフ構造を仮定して，ハッシュタグ間のユーザベース・単語ベースの AEMI (Augmented Expected Mutual Information) [9] による共起率とトピック分布の類似度を考慮した決定木を学習することで，2 つのハッシュタグ間の類義・対立・関連あり・関連なしという関係を推論することでハッシュタグを構造化した．しかし，この中では実際にツイートに対してハッシュタグを推薦するところまでは踏み込んでいない．

SNS の投稿について分類やタグの推定を行う場合には，投稿のテキスト特徴に加えて，投稿時間などのメタデータを利用できる場合が多い [10]．イベントに関連したタグの推定には，時間帯やユーザの嗜好などが有効な手がかりとして利用できると考えられるが，長期間にわたるイベントや，複数のイベントが同時間帯に行われている場合などではメタデータのみでは同定しきれないことが考えられるため，テキスト特徴の考慮は不

可欠である．本研究では，テキスト特徴のみを用いた手法に焦点を当て，より精度の高い手法を検討することによって，メタデータを用いる手法と組み合わせた際にも有効なアプローチを議論する．

3. 手　　法

本研究の主な目的は，SNS におけるタグごとの特徴を学習することで，ユーザのツイートからユーザがあるタグで表されるイベントに参加しているかどうかを検出することである．しかし，タグは必ずしも互いに独立に発生するとは限らず，表記上は異なるタグが同一の事象を表すことがしばしば起こる．そこで，できるだけ現実における事象と一対一に対応するようにタグをクラスタリングしておく必要がある．

本章では，本研究で提案するタグ文書コーパスの作成，タグ文書のクラスタリング，ユーザ文書に対するタグクラスタの推薦について具体的な手法を説明する．この章では，ある一定期間の SNS における全ての投稿から成る集合を D，D に出現する全てのタグから成る集合を $T(D)$ と表記する．

3.1 タグ文書コーパスの作成

SNS に現れる各タグの特徴量を得るために，ある特定のタグについてのタグ文書という仮想文書を考える．あるタグ $t \in T(D)$ についてのタグ文書 D_t は，$T(D)$ における投稿のうちタグ t を含む全てのテキストを結合したものとする．投稿には複数のタグを含むものもあるため，1 つの投稿が複数のタグ文書に含まれることもある．$T(D)$ に存在する各タグについてのタグ文書すべてを含むコーパスをタグ文書コーパス D_T とする．

3.2 タグ文書のクラスタリング

本研究の目的は，SNS にテキストを投稿しているユーザに対して，そのテキストの特徴から参加イベントを特定することである．一般的に考えれば，ある時間においてユーザが 1 つのイベントに参加していると仮定することで，ユーザに対して 1 つのイベントを推定すればいいことになる．しかし，本研究の提案する手法では，教師データに含まれるイベント関連ハッシュタグと，そのハッシュタグを含む投稿のテキストからそのイベントの特徴を学習し，ユーザが投稿したテキストに対して尤もらしいハッシュタグを推定することでユーザの参加イベントを間接的に特定する．

ここで問題になるのは，教師データに含まれるハッシュタグと現実世界におけるイベントの関係が一般的には多対多になっているということである．教師データとして扱うテキストが投稿された期間を広く取るほどこの特徴は強くなる．例えば，平日に毎日放送されるテレビのニュース番組に関するハッシュタグは，そのハッシュタグを含むテキストの特徴が毎日ニュースの内容によって変化するので，月曜日から金曜日までのデータでまとめて学習を行うと，1 つのイベントについての特徴が曖昧になってしまい，ユーザ文書からイベントを特定することが困難になってしまう．また，1 つのイベントに関して複数のハッシュタグが存在する場合は，ユーザ文書に対してハッシュタグを単独に推定した時に，同じイベントに関するハッシュタグが確率的に上位で推定されることが考えられるが，これは 1 つのイベントを推定したいという目的においては厄介な状況である．単純にこのような状況下で 1 つのイベントを推定する問題を解決するにはある程度複雑なモデルが必要と想定される．

本研究では，イベントに関するハッシュタグは複数あるが，あるハッシュタグからはイベントを一意に特定できるような期間に教師データの対象期間を狭めることで，ハッシュタグのクラスタリングの結果が現実世界における 1 つのイベントと結びつくような手法を提案することで，上記の問題を簡潔に解決することを目指す．

クラスタリングのタスクは迷惑メールの発見やニュース記事のジャンル推定など，クラスタ数のオーダが数十という程度であることが多い．しかし，毎日大量のテキストが投稿される SNS において，あるイベントを特定するためのハッシュタグ数が全体の異なりハッシュタグ数に対して数十分の一という状況は到底考えられない．そこで，約 1,000 の異なりハッシュタグ数に対してクラスタ数を 0.5–0.9 倍，すなわち数百から場合によっては数千というオーダで設定することで，互いに特徴の似たハッシュタグが 1 つのクラスタに統合され，それが結果的に現実世界における 1 つのイベントを指す現象が発生することを利用して，単独のハッシュタグの代わりにハッシュタグクラスタを推定することで一意なイベントを特定できるようにする．

具体的には，D_T が含むタグ文書 $D_{t \in T}$ を，k-means 法でクラスタリングする．各文書は TF-IDF ベクトルで表されるとする．

3.2.1 TF-IDF

TF-IDF とは，コーパス中の特定の文書における単語の重み付けの方法で, [11] で初めて提案された．この手法は，直感的には「ある文書におけるある単語が，他の文書には出現せずこの文書の中には頻出する」という場合にスコアが高くなるような式になっている．具体的には，コーパス中の文書数を N，コーパス中のある文書を d，d 中に出現するある単語を w，文書 d に単語 w が出現する回数を $\mathrm{Freq}(d, w)$，コーパス中で単語 w が出現する文書の数を DF^w とすると，ある文書 d における単語 w の TF-IDF スコアは式 3 で定義される：

$$\mathrm{TF}_d^w = \mathrm{Freq}(d, w) \tag{1}$$

$$\mathrm{IDF}^w = \log \frac{N}{\mathrm{DF}^w} + 1 \tag{2}$$

$$\mathrm{TF.IDF}_d^w = \mathrm{TF}_d^w \cdot \mathrm{IDF}^w \tag{3}$$

式 3 の左辺の表記から分かる通り，TF-IDF という重み付けは文書と単語の組に対して定義される．そのため，ある文書が持つ各単語に対して TF-IDF が計算できるので，コーパス中の各文書はその長さをコーパス中に出現する単語タイプ数とするベクトルで表現されることになる．文書の TF-IDF ベクトル中の各要素は，その文書中に出現する各単語が持つ TF-IDF スコアとなる．本研究では，前節で作成したタグ文書コーパスにおける各タグ文書の TF-IDF ベクトルを計算する．

3.2.2 k-means 法

k-means 法は，1967 年に [4] で提案されたクラスタリング手法である．各データ点はベクトルで表現されることを仮定している．簡単な説明は次の通りである．

（1） 各クラスタを代表する点（セントロイド (centroid) と呼ぶ）を，クラスタリング対象のデータ集合が存在する空間中にランダムに設定する

（2） 次の 2 つの手続きを，全てのクラスタのセントロイドの位置が収束するまで繰り返す

（a） 各データ点について，それが最も近いセントロイドを持つクラスタに属するようにする

（b） 各クラスタのセントロイドを，そのクラスタに属するデータ点の集合の重心になるように更新する

具体的なアルゴリズムは Algorithm 1 に示した．

Algorithm 1 k-means 法

k = クラスタ数
$D = n$ 次元データの集合
$C_{d \in D} = d$ が所属するクラスタの番号 $(1 \leqq C_d \leqq k)$
$\mathrm{Cent}_i = i$ 番目のクラスタのセントロイド
for $i = 1$ to k **do**
  Cent_i をランダムに初期化
end for
while 未収束 **do**
  for $d \in D$ **do**
    $C_d = \arg\min_k \mathrm{distance}(d, C_k)$
  end for
  for $i = 1$ to k **do**
    $\mathrm{Cent}_i = \sum D_{C_d=i} / n(D_{C_d=i})$
  end for
end while

3.3 ユーザ文書に対するタグクラスタの推薦

SNS におけるあるタグが現実世界におけるイベントを表していると仮定すると，SNS 中のユーザがそのイベントに参加しているかどうかは，イベントの時間中にユーザが投稿したテキストの集合（これを「ユーザ文書」と呼ぶ）がそのイベントを指すタグ文書を持つクラスタと近い距離にあるかどうかで判断できる．ユーザ文書に近いタグクラスタの選択方法は，次の 2 つが考えられる．

1 つ目は，各クラスタが持つセントロイドの位置とユーザ文書の距離を取る方法である．これは k-means 法によるクラスタリングの結果を直接使っていると言える．本稿ではこの手法を「重心法」と呼ぶことにする．

2 つ目は，ユーザ文書に最も近い k 個のタグ文書が属するクラスタの多数決で決める方法である．これは一般的に k-近傍法と呼ばれている．本稿ではこの手法を上記の「重心法」に対して「近傍法」と呼ぶことにする．

4. 実　　験

本研究の手法を評価するために，Twitter のツイートに対してハッシュタグを推定する実験を行った．

4.1 実験対象のハッシュタグと学習方法

対象とするハッシュタグは #precure, #giants, #図書館総合展の 3 つである．各ハッシュタグが対象とするイベントと，それが発生したとみなす日時は表 2 に示した．タグ文書コーパスに含む対象とするツイートは，各ハッシュタグについてのイベントが発生した日時を含む月曜日から日曜日までの 1 週間に投稿されたもののうち，10 文字以内のハッシュタグ文字列を含むツイート全てとする．具体的には，#precure と#giants は 2013 年 4 月 1 日から 2013 年 4 月 7 日までに投稿されたツイート，#図書館総合展 は 2012 年 11 月 19 日から 2012 年 11 月 25 日に投稿されたツイートを対象としてハッシュタグ文書コーパスを作成する．また，ハッシュタグ文書コーパスに含む対象のハッシュタグは，ツイートを収集する 1 週間の間に 100 回以上出現したものとする．その後，ハッシュタグ文書コーパスの各文書を TF-IDF ベクトルで表現したものを k-means 法でクラスタリングする．

4.2 ハッシュタグ推定対象のツイート

4.1 節で挙げた各ハッシュタグについて，そのハッシュタグに関するイベントの発生中に同一ユーザによって投稿された 5 件以上のツイートをすべて連結したものを 3.3 節におけるユーザ文書と見なしてハッシュタグの推定を行う．

4.3 実 験 内 容

本研究では次の 2 種類の実験を行う．

1 つ目は，最初から対象のハッシュタグが付けられていたツイート群に対してハッシュタグクラスタの推定を行うことで，最も正解率が高くなるクラスタ数を調査する実験である．本来 k-means 法は教師なしのクラスタリング手法のため文書分類には使われないが，本研究では先行研究を踏まえて，用意したハッシュタグ文書数に対してクラスタ数を大きくすれば上手く現実の概念に対応するクラスタリングが行えると仮定してこの方法で実験を行う．実験するクラスタ数は，ハッシュタグコーパス中のハッシュタグ文書の数に対して 0.5 倍から 0.9 倍まで 0.1 刻みの数で設定する．この実験における「正解」とは，ハッシュタグ推定対象のツイート群に対して推定したハッシュタグクラスタの中に，もともとそのツイート群に付けられていたハッシュタグが含まれている状態のことを指す．この実験では交差検定を行うが，得られるデータ量の関係から，#precure, #giants は 5 分割，#図書館総合展は 2 分割とする．

2 つ目は，1 つ目の実験で得られた最適なクラスタ数を用いて，イベントが行われた時間に投稿されたツイート群に対するハッシュタグ推定の精度と再現率を調査する実験である．この実験におけるハッシュタグ推定対象のツイート群は，あるイベントが行われていた時間帯のうちに同一ユーザによって 5 回以上投稿されたツイートの集合とし，それらのツイートを 1 つに結合したものをユーザ文書と見なす．このようなユーザ文書を，正しいハッシュタグが付けられていたものと何もハッシュタグが付けられていなかったもので同数用意する（用意する数は実験対象のハッシュタグごとに異なる）．正しいハッシュタグが付けられていたツイート群に対してはそのハッシュタグを含むハッシュタグクラスタが推定されるべきだが，ハッシュタグが付けられていなかったツイート群については，そのイベントに関係しかつそのイベントのハッシュタグが付けられていなかったツイート群というものが存在する．そのようなツイート群に対して実験対象のハッシュタグが推定された場合は正解と見

ハッシュタグ	対象とするイベント	イベントが発生した日時
#precure	テレビ朝日系列で放送されるアニメ『ドキドキ！プリキュア』の放送	2013 年 4 月 7 日 8 時 30 分-9 時 00 分
#giants	読売ジャイアンツが登場するテレビ野球中継	2013 年 4 月 3 日 18 時 00 分-21 時 00 分
#図書館総合展	第 14 回図書館総合展 1 日目	2012 年 11 月 20 日 10 時 00 分-18 時 00 分

表 2　実験に用いるハッシュタグの詳細

なす．また，実験対象のイベントに関係ないツイート群に対しては，実験対象のイベントのハッシュタグを含まないハッシュタグクラスタが推定された状態を正解と見なす．そこで，この実験における「精度」の母数は対象のハッシュタグを推定したユーザ文書の数,「再現率」の母数はその内容が対象のイベントに関連しているユーザ文書の数とし，どちらの場合もユーザ文書の内容が対象のイベントに関連しておりかつ対象のハッシュタグを推定していた場合を正解と見なす．

4.4　結　　果

対象のハッシュタグ付きのユーザ文書についてのハッシュタグ推定の結果は図 1,2,3 に示した．いずれも横軸が異なりハッシュタグ数に対して設定されたクラスタ数の比率，縦軸がユーザ文書に対して推定したハッシュタグクラスタの中に対象ハッシュタグが含まれていた割合である．図 1,2 のエラーバーは標準偏差を表している．図 3 については，2 分割の交差検定のため各テストケースを直接プロットした．ハッシュタグクラスタの推定方法に注目すると，3 つの対象ハッシュタグにおいて全てのクラスタ数比率で近傍法の精度が重心法と同じか上回っている．これは，クラスタの中心よりも個別のハッシュタグと比較した方が精度が高くなることを意味している．学習するクラスタ数の比率に注目すると，どちらの手法においても，訓練データに含まれる異なりハッシュタグ数の 0.6–0.7 倍にクラスタ数を設定して学習した場合がハッシュタグクラスタの推定精度が一番大きくなっている．0.8 や 0.9 など異なりハッシュタグ数に近い数にクラスタ数を設定すると，クラスタの中身が単独のハッシュタグに近い状態になることが考えられるが，この設定で精度が下がっていることは，本稿で今までに述べたとおりハッシュタグのクラスタリングが必要であることを示唆している．

上記で得られたクラスタ数比率を元に，対象ハッシュタグごとにそれに関連するイベントの内容が記述されているものとされていないものを対象としてハッシュタグの推定を行った実験の結果は表 3 の通りである．全体的な傾向として，あるイベントのハッシュタグを推定したユーザ文書はほぼ全てそのイベントに関連したテキストだったが，イベントに関連したテキストから成るユーザ文書に対する正しいハッシュタグクラスタの推定は重心法と近傍法の間で大きな差が出た．重心法によるハッシュタグクラスタの推定については，間違ったハッシュタグクラスタを推定した場合のほとんどが中身のハッシュタグが統一的でない曖昧なクラスタを推定していた．これは，そのクラスタの各ハッシュタグの特徴はユーザ文書とは程遠いが，そのようなハッシュタグの重心を取ると対象イベントのハッシュタグから成るクラスタの重心よりも近くなってしまうという現象に

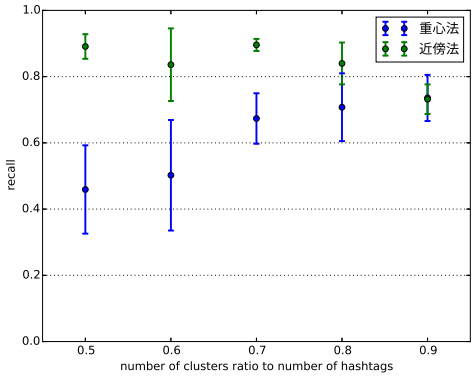

図 1　#precure の推定結果

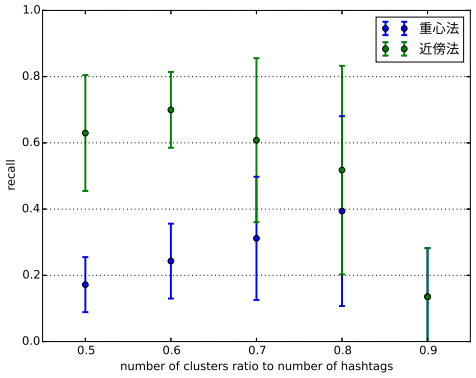

図 2　#giants の推定結果

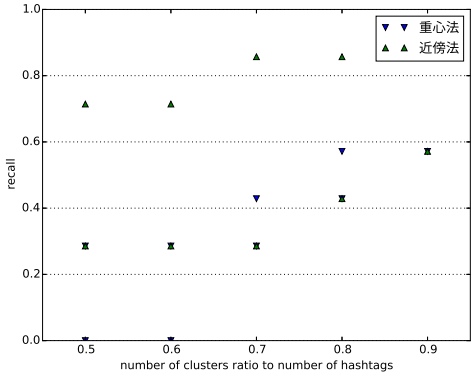

図 3　#図書館総合展の推定結果

よるものと考えられる．そのため，クラスタの重心と比較する重心法よりも，単独のハッシュタグと比較する近傍法のほうが精度が高かったと予想される．

#giants と #図書館総合展 については，いずれも近傍法で高い再現率を示しているものの， #precure と比べると劣る結果となった．2 つのハッシュタグについて共通する失敗例としては，ユーザ文書が極端に短かった場合に特徴を上手く抽出できず関係ないハッシュタグクラスタが推薦されたものがあった．#giants については，試合の観戦中に興奮して叫ぶような文字列が投稿された結果「#落ち着こう」,「#とりあえず叫ぼう」などが含まれるハッシュタグクラスタが推薦されたもの，対戦相手のチームについてのハッシュタグが含まれるハッシュタグクラスタが推薦されたもの，クラスタリングに失敗し#giants を含まないが読売ジャイアンツに関するハッシュタグが含まれるクラスタが推薦されたものが失敗例として見られた．#図書館総合展については，ユーザ文書が極端に長くかつ特徴的な語を多く含むユーザ文書に対して失敗している例が多くあったが，原因を特定することは出来なかった．

	#precure		#giants		#図書館総合展	
	重心法	近傍法	重心法	近傍法	重心法	近傍法
精度	1.000	0.993	1.000	1.000	1.000	1.000
再現率	0.290	0.816	0.100	0.520	0.143	0.500
F 値	0.450	0.896	0.182	0.684	0.250	0.667

表 3　各ハッシュタグにおける精度と再現率

精度と再現率の実験の過程で得られたハッシュタグクラスタの例は表 4 に示した．#giants を含むハッシュタグクラスタには #baystars が含まれているが，これは訓練データの取得対象とした期間中に読売ジャイアンツと横浜ベイスターズの試合が行われたのが理由として考えられる．

対象ハッシュタグ	対象ハッシュタグを含むハッシュタグクラスタ
#precure	[precure, ドキプリ, nitiasa]
#giants	[giants, mlbjp, baystars, Giants]
#図書館総合展	[図書館総合展]
その他の例	[高校野球, kokoyakyu, 甲子園]
	[とびだせどうぶつの森, どうぶつの森, とび森]
	[エイプリルフール, 4 月 1 日, エイプリールフール, エープリルフール, 嘘, 四月馬鹿]

表 4　得られたハッシュタグクラスタの例

5. ま と め

本研究では，SNS に投稿されたイベント期間中のテキストを用いて，そのテキストの集合に対してタグを推定することでそのユーザがイベントに参加しているかどうかを推定する手法を提案した．Twitter のツイートデータとハッシュタグで実験を行った結果，クラスタの中身が単独のハッシュタグに近い状態となる高いクラスタ数比率よりも，ある程度の数のハッシュタグがクラスタの中に存在する 0.6–0.7 といった中程度のクラスタ数比率の方が精度が高くなった．この結果は本研究で事前にハッシュタグのクラスタリングを必要とした理由であるハッシュタグとイベントの多対一の関係性が現れた結果と言える．精度と再現率を示す実験においては，あるイベントに関係ないツイート集合に対して対象ハッシュタグを含むハッシュタグクラスタを推薦することがほぼ無く，イベントに関連する内容を含むツイート集合に対しても高い確率で正しいハッシュタグを推定できていたことから，あるイベントについて検索した時に関係ないツイートを取得する確率はかなり低いと言える．

また，現存する SNS のほぼ全てがタグという仕組みを有しているため，Twitter に限らずそれら全てのサービスに対して本手法が適用できるのは本研究の強みである．

一方で，当然全てのイベントに対してハッシュタグが定義されているとは限らない．今回の実験でも対象のハッシュタグは天下り的に与えたものである．今後は現実で定期的に発生するイベントに関するハッシュタグや，事前に定義されたハッシュタグが無いようなイベント（特に突発的に発生するものなどはそうである）における関連語を，時間枠ごとに特異値的に現れた単語として定義することで自動的にイベントに関連する特徴語を取得するような研究が必要と考えられる．

謝　　辞

本研究の一部は，JSPS 科研費（課題番号 25280110, 25540159）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 (Research Projects of Facalty of Library, Information and Media Science) の助成によって行われた．

文　　献

[1] 伊川洋平 and 村上明子. Twitter におけるイベントモニタリングのためのノイズ除去. In **第 *13* 回日本データベース学会年次大会**, 2015.

[2] Oren Tsur, Adi Littman, and Ari Rappoport. Efficient Clustering of Short Messages into General Domains. In *International Conference on Weblogs and Social Media (ICWSM)*, 2013.

[3] Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704, 2011.

[4] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[5] Dolan Antenucci, Gregory Handy, Akshay Modi, and Miller Tinkerhess. Classification of tweets via clustering of hashtags eecs 545 final project, fall, 2011. Technical report, 2011.

[6] Fréderic Godin and V Slavkovikj. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 593–596, 2013.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[8] 木村輔 and 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. In **第 *12* 回日本データベース学会年次大会**, 2014.

[9] Philip K Chan. A non-invasive learning approach to building web user pro les 1 Introduction 2 Page Interest Estimator (PIE). *KDD-99 Workshop on Web Usage Analysis and User Profiling*, 1999.

[10] Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 999–1008, New York, NY, USA, 2014. ACM.

[11] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.