

Twitter ハッシュタグを用いた類似イベント検索

本岡 亮[†] 湯本 高行^{††} 新居 学^{††} 高橋 豊^{††} 角谷 和俊^{†††}

[†] 兵庫県立大学工学部 〒 671-2280 兵庫県姫路市書写 2167

^{††} 兵庫県立大学大学院工学研究科 〒 671-2280 兵庫県姫路市書写 2167

^{†††} 兵庫県立大学環境人間学部 〒 670-0092 兵庫県姫路市新在家本町 1-1-12

E-mail: [†]eo07g112@steng.u-hyogo.ac.jp, ^{††}{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,

^{†††}sumiya@shse.u-hyogo.ac.jp

あらまし 既知のイベントと類似した未知のイベントを発見することは一般に困難である。そこで本研究では、Twitter で使用されるハッシュタグと、そのハッシュタグが示すイベントの対応関係に基づいて、与えられたハッシュタグに対する類似イベントを示すハッシュタグを発見する手法を提案する。具体的には、与えられたハッシュタグを使用しているユーザ集合と、そのユーザ集合が使用しているハッシュタグ集合より、与えられたハッシュタグと取得した各ハッシュタグ間の類似を判定する。また、取得した各ハッシュタグが使用された投稿時間の分布により、ハッシュタグが示す内容のイベント性の判定を行うことによって、与えられたハッシュタグが示すイベントの類似イベントを示すハッシュタグを発見する。

キーワード Twitter, ハッシュタグ, イベント検索

A Similar Event Search System Using Hashtag of Twitter

Ryo MOTOOKA[†], Takayuki YUMOTO^{††}, Manabu NII^{††}, Yutaka TAKAHASHI^{††}, and Kazutoshi SUMIYA^{†††}

[†] School of Engineering, University of Hyogo. 2167 Shosha, Himeji, Hyogo, 671-2280, Japan

^{††} Graduate School of Engineering, University of Hyogo. 2167 Shosha, Himeji, Hyogo, 671-2280, Japan

^{†††} School of Human Science and Environment, University of Hyogo. 1-1-12 Shinzaikake-honcho, Himeji, Hyogo 670-0092, Japan

E-mail: [†]eo07g112@steng.u-hyogo.ac.jp, ^{††}{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,

^{†††}sumiya@shse.u-hyogo.ac.jp

1. はじめに

インターネットを用いた情報検索は我々の生活に深く関わっている。その用途は様々であるが、その中でも自分の興味のあることを調べたり、知識の幅を広げたりするために使われていることは多いと考えられる。

しかし、Web を用いて自分は知らないが興味のある事柄を探そうとする場合、そもそも存在を知らず検索クエリをどのように設定したら良いのか分からないため、発見は非常に困難である。その中でも、既知のイベントと類似したイベントを発見するためには、多くの情報を探索・閲覧し、必要な情報のみを取捨選択する必要があるためユーザの負担が非常に増えてしまう。

一方、近年 Social Networking Service, ブログといったソーシャルメディアが普及してきている。これらのソーシャルメ

ディアには、使用ユーザの実生活に関する情報が多く発信されている。この実生活の中には使用ユーザのイベントに関する興味、感想等も含まれているためソーシャルメディアはイベントに関する情報収集に適していると考えられる。その中でも Twitter [1] のユーザ数は爆発的に増加しており、様々な種類の情報を容易に発見可能である。さらに Twitter にはハッシュタグというタグが存在し、タグ検索が可能である。Twitter はソーシャルメディアの中のマイクロブログに属しており、その投稿の事を tweet と呼ぶ。ハッシュタグは Twitter の投稿時に使用されるタグである。

本研究では、Twitter に使用されるハッシュタグに注目して、既知のイベントを示すハッシュタグからその類似イベントを示すハッシュタグを検索・発見することを目的とする。検索手法としては、ユーザ集合より取得した各ハッシュタグが使用され

た tweet の時間経過による盛り上がり方から各ハッシュタグのイベント性を判定する。更に入力ハッシュタグを使用しているユーザ集合と、そのユーザ集合の tweet に使用されているハッシュタグ集合から、入力ハッシュタグと取得した各ハッシュタグの類似性を判定することによって、入力ハッシュタグが示すイベントと類似したイベントを示すハッシュタグを発見する。

本稿では、入力ハッシュタグを使用しているユーザから取得した各ハッシュタグのイベント性及び類似性を判定する手法について述べる。以下、2 章では関連研究、3 章では類似イベントを示すハッシュタグの検索手法、4 章では評価実験について述べる。

2. 関連研究

田中ら [2] は Twitter で使用されるハッシュタグに注目し、ハッシュタグを含む tweet を時間経過に沿ってクラスタリングすることで、tweet 集合の時間経過と共に変化する内容抽出を行っている。ハッシュタグをイベント発見のためのものではなく、複数の tweet からなる仮想的な場として捉えている点が本研究とは異なる。

藤坂ら [3] は位置情報付きの tweet を取得し、発信したユーザの行動の状態に注目することによって、地域イベントを発見する手法を提案している。ユーザから類似イベントを探すのではなく、人間の行動に基づいた地域イベントを探すという点で本研究と異なっている。

岩木ら [4] は、マイクロブログにおける有用な記事の発見支援を行っている。著者に興味があるか・内容が万人向けであるか、という 2 通りの側面から記事が有用であるか判定する。著者への興味を、ブロガーの友達概念のリンク構造や記事に対するコメント等から抽出している。本研究では、類似イベントを推薦するため Twitter におけるフォロー関係や、返信等は考慮せずに入力ハッシュタグと関連したハッシュタグを出力することが可能となっている。

3. ハッシュタグ検索手法

3.1 Twitter・ハッシュタグ

Twitter はマイクロブログという Web 上のサービスの 1 つであり、1 投稿あたりの文字数が 140 までと制限されている。用途は様々だが、個人的な呟きや情報収集などを主流に使用されている。新聞社や放送局の中にはアカウントを作り Twitter 上にニュースを流している団体もあり、イベントなどの主催者がオフィシャルアカウントとして更新情報を発信することもある。

ハッシュタグは tweet 中に “#worldcup” のように “#英数字列” の形で使用される。基本的に同じハッシュタグが使用されている tweet 集合は同じイベントやカテゴリに対して言及している。例えば、図 1 に示すように “#worldcup” というハッシュタグが使用された tweet はワールドカップについて言及している。

3.2 イベント・類似の定義

本稿で扱うイベントの定義及び類似の定義をする。まず、以下の 2 つの条件を両方満たす事象をイベントと定義する。

- 開催されない期間に比べて開催期間が十分に短い事象

試合終了。1-0でスペインの優勝 [#nhk_fifaworldcup #worldcup](http://nhk.jp/Ewm64)
6:07 AM Jul 12th NHKから

図 1 ハッシュタグ

- 開催日時、場所等が事前に告知してある事象

連続テレビドラマなど開催期間の間隔が短い事象や、「Jリーグ」のような毎年の約 8 割の期間に渡って開催されている事象は本稿ではイベントとみなさない。また、突発的なニュースや天災等、起こる前に何らかの予兆がない事象もイベントとみなさない。

類似イベントの定義は、入力ハッシュタグが示すイベントと同カテゴリに属しているイベントとする。例えば “#worldcup” が入力ハッシュタグならば、サッカーというカテゴリに属するイベントを類似イベントとする。

3.3 提案手法の概要

本研究では既知のイベントを示すハッシュタグを入力とし、Twitter 上で入力ハッシュタグを使用しているユーザを取得する。その各ユーザの tweet を取得し、その tweet 中に使用されているハッシュタグを取得・解析することによって、入力ハッシュタグが示すイベントと類似したイベントを示すハッシュタグを発見する。

処理手順を以下に示す。

(1) 入力ハッシュタグを使用しているユーザを Twitter から l 人分取得し、そのユーザ集合が使用しているハッシュタグを取得

(2) (1) で取得したハッシュタグ集合に対してイベント性判定を行う

(3) 類似性に基づき、ランク付けを行う

(2) の処理は、ハッシュタグが示している内容のイベント性を判定するために集中度という尺度を用いたフィルタリングと突発事象除去のためのフィルタリングである。集中度はハッシュタグが使用された tweet の投稿時間の分布によって本研究で定義した開催期間に関する条件を、突発事象は投稿時間が集中している期間の直前の期間によって事前情報の有無に関する条件に対する判定を行う。

(3) の処理は、類似を判定するために補助となる認知度という尺度を用いてフィルタリングを行った後に、入力ハッシュタグと取得した各ハッシュタグ間の類似度を定義し、ランク付けを行う。認知度を用いたフィルタリングは、有名なハッシュタグを類似イベントを示すハッシュタグである、と誤判定することを防ぐために行う。類似度によるランク付けは、入力・取得ハッシュタグを使用しているユーザ集合より判定する。

最後に、ランク付けを行ったハッシュタグ集合の上位 n 件を類似イベントを示すハッシュタグとして出力する。

提案手法の全体像を図 2 に示す。

3.4 イベント判定手法

ハッシュタグが本研究で定義したイベントを示しているか、

図2 処理の流れ

開催期間及び事前情報の2つの観点から判定を行う。期間からイベントであるか判定するための集中度を用いたフィルタリング関数を σ_{Con} 、突発事象を除去するためのフィルタリング関数を σ_{Pre} 、取得ハッシュタグ集合を H とすると、イベント判定のためのフィルタリング関数 $\sigma_{Eve}(H)$ を(1)式のように定義する。

$$\sigma_{Eve}(H) = \sigma_{Con}(\sigma_{Pre}(H)) \quad (1)$$

3.4.1 集中度によるフィルタリング

ユーザ集合から取得したハッシュタグ h' が示している内容がイベントを示すかどうかを判断するための尺度として集中度を定義し、それを用いてフィルタリングを行う。集中度によるフィルタリングは、イベントの開催回数及び開催期間を判定するために行う。

取得ハッシュタグ h' が使用された tweet を m_{con} 個取得し、その投稿された時間情報を T 日ずつに tweet 数を分割する。そして、その中の最大値で各期間の使用回数を正規化したヒストグラムを作成する。各期間の最大値と tweet 数を正規化した値の差分の総和を、イベントの開催回数及び開催期間を判定するための集中度と定義する。また、集中度算出のために取得するハッシュタグが投稿された期間は2010年1月から2011年1月とした。

各分割した期間 t_i でのハッシュタグ h' の使用回数を $UseCount(h', t_i)$ 、使用数が最も多い期間を t_{max} とすると、分割した期間の中の最大使用回数を $UseCount(h', t_{max})$ となり、取得ハッシュタグ h' の集中度 $Con(h')$ を(2)のように定義する。

$$Con(h') = \sum_{i=0}^n \left(1 - \frac{UseCount(h', t_i)}{UseCount(h', t_{max})} \right) \quad (2)$$

いずれのハッシュタグでも tweet の取得数は m_{con} 個なので最大値で正規化することによって、ヒストグラムの値の総和に差が発生する。イベントを示すハッシュタグはイベント開催期間、またはその前後の期間に集中して使用され、その他の期間ではほとんど使用されていないので、最大値で正規化することによってイベント開催期間とその前後の期間の値だけが残し、ヒストグラムの値の総和は低くなり集中度は高くなる。

逆にイベントを示さないハッシュタグは時期に関係なく使用されるため、ヒストグラムの総和は高くなり集中度は低くなる。

この集中度が閾値 θ_{eve} よりも高い値を示せば、取得ハッシュタグ h' が示す内容がイベントであると判定する。本研究では集中度が閾値 θ_{eve} よりも低い値を示すハッシュタグはイベントを示していないとみなす。

例として、イベントである「ルビー会議」を示すハッシュタグ “#rubykaigi” と地名である「東京」を示すハッシュタグ “#tokyo” それぞれを使用した tweet を250件ずつ取得し、月

毎に分割し最大値で正規化した結果の2010年の4月から12月を抜粋したヒストグラムを図3、図4にそれぞれ示す。

図3では8月の投稿数が多いことが確認できる。このことから「ルビー会議」というイベントは8月に行われたと推測することができる。逆に図4からは“#tokyo”というハッシュタグが時期に関係なく使用されていることが確認できる。このような日常的に使用されているハッシュタグはイベントを示していないと推測することができる。

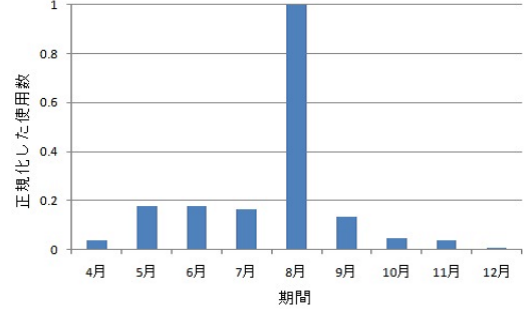


図3 #rubykaigi の月別使用傾向

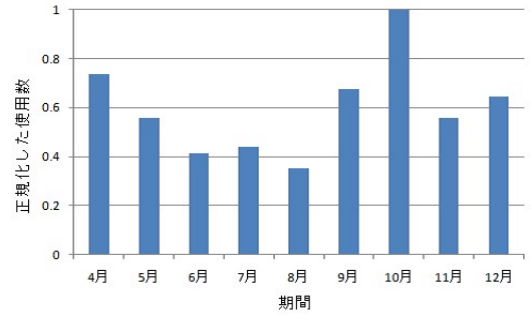


図4 #tokyo の月別使用傾向

ユーザ集合から取得したハッシュタグ集合を H 、イベント判定のための閾値を θ_{eve} とすると、フィルタリング関数を(3)式のように定義できる。

$$\sigma_{Con}(H) = \left\{ h' \mid h' \in H, Con(h') > \theta_{eve} \right\} \quad (3)$$

3.4.2 突発事象の除去手法

取得ハッシュタグ h' が示す事象が、以前から計画されていたかどうかを確認するためのフィルタリングを行う。取得ハッシュタグ h' が使用された tweet の期間内使用最大数を基準として、 k 個前までの期間に tweet 数が0である期間の数が閾値 θ_{pre} 以下の場合、ハッシュタグ h' は事前に計画されたイベントを示すと判定する。ハッシュタグ h' が使用された tweet 数が最大となる期間を t_{max} とし、そこから k 個前までの期間において、tweet 数が0の期間の数を $CountZero(h', t_{max}, k)$ 、その閾値を θ_{pre} 、ユーザ集合から取得したハッシュタグ集合を H とするとフィルタリング関数を(4)式のように定義する。

$$\sigma_{Pre}(H) = \left\{ h' \mid h' \in H, CountZero(h', t_{max}, k) < \theta_{pre} \right\} \quad (4)$$

3.5 類似判定手法

ユーザ集合より取得した各ハッシュタグが示す事象が、入力ハッシュタグが示す事象と類似しているか判定するために、類似度、認知度という2つの尺度を用いる。類似度は、入力ハッシュタグと取得ハッシュタグそれぞれを使用しているユーザ集合によって、ハッシュタグ同士の類似を判定する尺度である。また、有名なハッシュタグは多くのユーザに使用されているため類似度は高くなる傾向にあり、類似していると判定されてしまう。このことを防ぐために類似判定の補助として、そのハッシュタグが Web 上でどの程度認知されているかを表す認知度という尺度によるフィルタリングを行う。類似度算出の前に認知度を用いたフィルタリングを行うので、先に認知度について記す。

3.5.1 認知度を用いたフィルタリング

有名なハッシュタグは、多くのユーザに使用されているので、類似していると判定されてしまう可能性がある。そのような誤判定を防ぐために、ユーザ集合から取得したハッシュタグ h' がどの程度 Twitter 上で認知されているかを表す尺度として認知度を定義する。tweet 数が多いハッシュタグほど認知度が高くなる。

また、ほとんど認知されていないハッシュタグは個人的な使用方法のものが多くイベントを示していない場合が多い。そのため認知度の上限と下限を設定し、フィルタリングを行う。

本研究では、各ハッシュタグの tweet 数を確認するために Twitter 上の tweet のデータを長期に渡って保持している検索エンジンの1つである Topsy [5] を用いる。Topsy は検索対象として、Web, Photo, Tweets などを設定することができ、またそれぞれの対象について、今日、1週間以内、1ヶ月以内、全時間等の期間の選択が可能である。取得ハッシュタグ h' の Topsy における Tweets を対象にした全時間での検索結果件数を認知度とする。

$Hit(h')$ を Topsy での取得ハッシュタグ h' を使用した tweet の検索結果件数、検索結果件数 $Hit(h')$ の上限を θ_{upper} 、下限を θ_{lower} 、イベント性判定後のハッシュタグ集合を H' として、認知度によるフィルタリング関数を (5) 式のように定義する。

$$\sigma_{Rec}(H') = \left\{ h' \mid h' \in H', \theta_{lower} \leq Hit(h') \leq \theta_{upper} \right\} \quad (5)$$

3.5.2 類似度によるランキング

ユーザ集合から取得したハッシュタグ h' が示すイベントが、入力ハッシュタグ h が示すイベントとどの程度類似しているかを表す尺度として類似度を定義する。取得した入力ハッシュタグ h' の使用ユーザの中で、取得ハッシュタグ h を使用しているユーザの割合を類似度と定義する。Twitter で入力ハッシュタグ h を使用しているユーザの中から l 人取得したユーザ集合を $User(h)$ とする。そのユーザ集合の中で取得ハッシュタグ h' を使用しているユーザ集合 $User(h'; h)$ を (6) 式のように表す。

$$User(h'; h) = \{ u \mid u \text{ が } h' \text{ を使用している}, u \in User(h) \} \quad (6)$$

入力ハッシュタグ h に対する取得ハッシュタグ h' の類似度

$Sim(h, h')$ を (7) 式のように定義する。

$$Sim(h, h') = \frac{|User(h'; h)|}{|User(h)|} \quad (7)$$

3つの尺度の内、類似度以外の2つはフィルタとしての働きをするので、取得ハッシュタグのランキングは類似度によってのみ行う。

ユーザ集合より取得したハッシュタグ集合 H に対してイベント判定と類似判定のフィルタリング関数 σ_{Rec} と σ_{Con} を適用したハッシュタグ集合を H'' とすると、(8) 式のように表される。

$$H'' = \sigma_{Rec}(\sigma_{Eve}(H)) \quad (8)$$

イベント性・類似性判定後のハッシュタグ集合 H'' を、類似度 $Sim(h, h')$ でランキングした上位 n 位までを出力する。

4. 評価実験

提案手法が有用であるかどうか確認するために以下に示す2つの観点から実験を行った。

- ハッシュタグの出力傾向
- 時間経過に伴う出力の変化

実験の際の各パラメータを表1に示す。

2つの実験で出力されたハッシュタグが類似イベントを示しているものかどうかを Web を用いて人手で判断した。その際に、Google [6] を用いた検索以外にハッシュタグクラウド [7], hashtagsjp [8] の2つのサイトを使用した。これら2つのサイトは日本の Twitter のハッシュタグの情報をまとめ、ハッシュタグをより活用できるように支援するサイトであり、誰でもハッシュタグの情報を編集できるようになっている。

4.1 出力傾向確認のための実験

複数の異なるハッシュタグを入力とした際、それぞれの出力ハッシュタグが類似イベントを示した割合を確認する。また、そのハッシュタグがどのような事象を示しているかを確認するために実験を行った。

入力ハッシュタグは表2に示す5つのものを用いた。

表2 実験1の入力ハッシュタグ

入力ハッシュタグ	内容
#fujirock	FUJI ROCK FESTIVAL
#summersonic	SUMMER SONIC
#rubykaigi	Ruby 会議
#webdbf2010	WebDB Forum 2010
#afc2011	アジアカップ 2011

表2の各ハッシュタグを入力した際の出力ハッシュタグが類似イベントを示した適合率を表3に示す。また、各ハッシュタグの出力結果を表4から表8に示す。表4から表8の正解の項目は、ハッシュタグの示す内容を人手で確認し、類似イベントを示している内容であれば“○”，そうでなければ“×”としている。また、表5の“#summersonic”を入力した場合、イベント判定及び類似判定によるフィルタリングによって残ったハッシュタグが10件に満たなかったため、得られた8件について

表 1 実験パラメータ

パラメータ	意味	値
l	ユーザの取得人数	50
m_{user}	ユーザからの取得 tweet 数	200
m_{con}	集中度算出のための取得 tweet 数	250
T	(2) 式における tweet の投稿数分割日数	10
θ_{eve}	(2) 式におけるイベント判定の閾値	36.0
k	(4) 式における tweet 最大期間直前の確認する期間数	3
θ_{pre}	(4) 式における tweet 数 0 の期間数の閾値	2
θ_{upper}	(5) 式における認知度の下限	100
θ_{lower}	(5) 式における認知度の上限	10000
n	出力件数	10

評価した。表 8 では、入力ハッシュタグと同じ「アジアカップ」を示すハッシュタグが出力された。同カテゴリのイベントだが、類似イベントではなく同一イベントを示しているため、不正解とした。

表 3 出力ハッシュタグが類似イベントを示した適合率

入力ハッシュタグ	適合率
#fujirock	0.50
#summersonic	0.25
#rubykaigi	0.80
#webdbf2010	0.50
#afc2011	0.20

表 6 入力#rubykaigi の場合の出力結果

ハッシュタグ名	内容	正解
#atbc	Agile Team Building Conference	○
#devsumi	Developers Summit	○
#cloudstock	Cloudstock	○
#tork03	東京 Ruby 会議 03	○
#97prog_ja	『プログラマが知るべき 97 のこと』 トークセッション	○
#rubyconf	RubyConf 2010	○
#sprk03	札幌 Ruby 会議 03	○
#spcamp	セキュリティ&プログラミングキャンプ	○
#potpub	ポットチャンネル (USTREAM 番組)	×
#oblove	オブラブ (グループ)	×

表 4 入力#fujirock の場合の出力結果

ハッシュタグ名	内容	正解
#cdj	COUNTDOWN JAPAN 10/11	○
#bigbeachfes	bigbeachfes	○
#rsr11	RISING SUN ROCK FESTIVAL 2011	○
#asagiri	ASAGIRI JAM	○
#arabaki	ARABAKI ROCK FEST	○
#zacjapan	ザッケローニ JAPAN	×
#primalscreeam	Primal Scream(アーティスト)	×
#libertines	The Libertines(アーティスト)	×
#bs1	NHK BS 1	×
#whitelies	White Lies(アーティスト)	×

表 7 入力#webdbf2010 の場合の出力結果

ハッシュタグ名	内容	正解
#pyconjp	PyCon mini JP	○
#rrds3	第 3 回楽天研究開発シンポジウム	○
#interbee	Inter BEE 2010	○
#msc2010	Microsoft Conference Japan Tour 2010	○
#scis2011	暗号と情報セキュリティシンポジウム	○
#sigmus	情報処理学会 音楽情報科学研究会	×
#jtf2010	JFT 翻訳祭	×
#daihyou	日本代表	×
#pi10	式根島冒険ツアー	×
#riasa	日本名門酒会「立春朝搾り」	×

表 5 入力#summersonic の場合の出力結果

ハッシュタグ名	内容	正解
#quasimode_ust	クオシモードライブ配信	○
#wh2010	WORLD HAPPINESS 2010	○
#cbc1053	CBC ラジオ 1053KHz	×
#13assassins_app	十三人の刺客 (iPhoneApp)	×
#levele	レベル E	×
#loomlab	LoomLab(アーティスト)	×
#sb21cp	ソフトバンクキャンペーン	×
#lvup	内容不明	×

表 3 より、情報関連の学会を示すハッシュタグである“#rubykaigi”，“#webdbf2010”を入力とした場合、出力ハッ

ッシュタグが類似イベントを示した適合率は高い値を示した。しかし、音楽関連のフェスティバルを示すハッシュタグである“#fujirock”，“#summersonic”，またスポーツの大会を示す“#afc2010”を入力とした場合はあまり良い結果を得ることが出来なかった。

表 4 から表 8 より、出力された類似イベントを示さないハッシュタグにいくつかの傾向が見られた。以下にその内容を述べる。

1 つ目は、表 4、表 8 の“#zacjapan”のように直接イベントを示してはいないが、イベントに関係のある内容を示すハッシュタグである。“#zacjapan”はサッカー日本代表を示すハッシュタグである。実験を行った日付近で「アジアカップ」というサッカーの国際大会が行われており、その大会に日本代表が出

表 8 入力#afc2011 の場合の出力結果

ハッシュタグ名	内容	正解
#koukou_fb	全国高校サッカー選手権	○
#90th_jyoho	第 90 回天皇杯全日本サッカー選手権大会	○
#asiacup2011	アジアカップ 2011	×
#asia_cup	アジアカップ 2011	×
#zacjapan	ザッケローニ JAPAN	×
#jfa2011	日本サッカー協会 2011	×
#bs1	NHK BS 1	×
#daihyou	日本代表	×
#nhk_asian_cup	アジアカップ 2011(NHK)	×
#hokoten	ホコテン再開	×

場していた。そのため、「アジアカップ」を示す“#afc2011”と“#zacjapan”のハッシュタグが同じ tweet に一緒に使用されることが多かったため、“#zacjapan”がイベントを示すハッシュタグであると判定され出力された。

2 つ目は、表 4 の“#primalscream”のようにアーティストを示すハッシュタグである。アーティストが行うライブイベントに固有のハッシュタグが存在しない場合、イベントに関係した tweet にアーティストそのものを示すハッシュタグを使用することが多い。このようなハッシュタグは、アーティストのニュース全般にも使用されるため、複数の事象を示すことになる。そのため、本研究で設定した尺度を用いたフィルタリングを適用しても良い結果が得られなかった。

3 つ目は、表 6 の“#oblove”のように団体を示すハッシュタグである。団体が年に数回イベントを開催しているが、そのイベント全てに同じハッシュタグが使用されているため、アーティストの場合と同様にハッシュタグが複数の事象を示してしまうため、イベントを示すと誤判定された。

また、出力されたハッシュタグ全体の問題を以下に記す。

表 8 の“#jfa2011”のようにカテゴリは一致しているがイベントを示さないハッシュタグが出力された。本来ならば、イベント判定の段階で出力候補から除外されるはずのハッシュタグが残った。イベント判定フィルタには改善の必要がある。

また、表 8 の入力ハッシュタグである“#afc2011”及び出力ハッシュタグである“#asiacup2011”と“#asia_cup”は同じ「アジアカップ」というイベントを示していることが確認できた。今回の場合、入力ハッシュタグと同じ事象を示すハッシュタグであるという点を除くと、イベントを示すハッシュタグを出力できているが、複数ハッシュタグが同じイベントでない事象を示していた場合、各ハッシュタグが使用された tweet の情報が分散し、集中度及び認知度を正確に算出できないため、イベントを示していないハッシュタグがイベントを示していると誤判定して出力してしまうことも考えられる。そのため、同じ事象を示すハッシュタグは可能な限り統一する必要があるが、現時点ではハッシュタグが示す事象を網羅できないため難しいと考えられる。

4.2 時間経過と出力変化の確認のための実験

Twitter 上のデータは日々変化しているので、時間経過による出力ハッシュタグが類似イベントを示した適合率の変化を確

認するために 2 種類の入力ハッシュタグで 5 日に渡って実験を行った。入力ハッシュタグを音楽系フェスティバルである「フジロックフェスティバル」を示す“#fujirock”，情報関連の学会である「ルビー会議」を示す“#rubykaigi”を用いた。

実験において、出力ハッシュタグが類似イベントを示した適合率を表 9 に示す。入力ハッシュタグが“#rubykaigi”の場合、

表 9 出力ハッシュタグが類似イベントを示した適合率

	1 回目	2 回目	3 回目	4 回目	5 回目	平均
#fujirock	0.30	0.30	0.50	0.60	0.50	0.44
#rubykaigi	0.70	0.90	0.80	0.80	0.80	0.80

表 9 に示すように適合率は変動するものの安定した結果が得られた。しかし、“#fujirock”の場合は適合率が“#rubykaigi”の場合より低くなり、適合率の値の最大値と最小値の差も大きくなった。

出力ハッシュタグが類似イベントを示した適合率が“#rubykaigi”のものより低くなった理由として、それぞれのハッシュタグを使用しているユーザ層の違いが挙げられる。それぞれのハッシュタグを使用しているユーザの tweet を人手で確認したところ、“#fujirock”を使用しているユーザ層は“#rubykaigi”を使用しているユーザ層よりも幅広いカテゴリに関して日常的に tweet していた。“#fujirock”を入力とした場合、類似イベントは音楽関連の催し物としたが、出力されたハッシュタグはお笑いイベントやスポーツイベントを示すものも存在し、これらのハッシュタグを不正解と判定したため適合率は低くなったと考えられる。

適合率の値の最大値と最小値の差が大きくなった理由も“#fujirock”を使用しているユーザに関係があると考えられる。“#rubykaigi”を使用しているユーザ層よりも、“#fujirock”を使用しているユーザ層の方が日常的な tweet 数が多いので、各ユーザから取得した tweet、 m_{user} 個に使用されているハッシュタグの毎回の変化が大きかったためだと考えられる。

5. おわりに

本稿では、Twitter に使用されているハッシュタグに注目し、ハッシュタグのイベント性、入力ハッシュタグと取得ハッシュタグとの類似性を判定することによって、入力ハッシュタグが示すイベントと類似したイベントを示すハッシュタグを発見する手法について述べた。

評価実験より、入力ハッシュタグを使用しているユーザ層によって精度は変動するものの、類似イベントを示すハッシュタグを発見することができた。しかし、イベントを示すハッシュタグと頻繁に併記されるハッシュタグや、複数事象を示すハッシュタグに関してはうまく尺度が働かず、類似イベントを示さないハッシュタグも出力されたので、これを改善する手法の考案が今後の課題である。

謝 辞

本研究の一部は、平成 22 年度科研費基盤研究 (B)(2)「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」

(課題番号：20300039) によるものです。ここに記して謝意を表すものとします。

文 献

- [1] Twitter. <http://twitter.com/>.
- [2] 田中克明, 堀浩一. Twitter ハッシュタグに基づくツイート群からの変化抽出. 電子情報通信学会 第二種研究会資料, pp.31-32, 2010.
- [3] 藤坂達也, 李龍, 角谷和俊. 実空間マイクロブログ分析による地域イベントの影響範囲推定. DEIM Forum 2010 D7-4(2009).
- [4] 岩木 祐輔, アダム ヤトフト, 田中 克己: マイクロブログにおける有用な記事の発見支援, DEIM Forum 2009 A6-6, (2009).
- [5] Topsy. <http://topsy.com/>.
- [6] Google. <http://www.google.co.jp/>.
- [7] ハッシュタグクラウド. <http://hashtagcloud.net/>.
- [8] hashtagsjp. <http://hashtagsjp.appspot.com/>.