Twitterにおける話題語の抽出と周期に基づく分類

佐々木 謙太朗　田村 一樹　吉川 大弘　古橋 武
名古屋大学

1　はじめに

近年急速に普及し，注目を集めている情報源としてTwitter[1]に代表されるマイクロブログがある．Twitterでは，ユーザーはツイートと呼ばれる140文字以内の文を投稿し，それを他のユーザーがフォローすることで購読するという形のサービスが提供されている．140文字以内という制限により，ユーザーは気軽に情報を発信でき，またそれによりTwitterは速報性，リアルタイム性が極めて高い情報源となっている．そのため，Twitterから有益な情報を得ようとする研究が近年盛んに行われている[2][3].

Twitterのような文書ストリームでは，時間を問わず文書情報が送られてくるが，ある時から特定の話題に関する記述が急激に増加するような現象が起こることがある．このような現象をバーストと呼ぶ[4]. Twitterにおいては，実世界で何らかのイベントが起きた際，そのイベントに関連するツイートのバーストが即時的に起きることがある．同様の現象は一般のブログにおいても起こるが，Twitterではより顕著である．例えば地震のような広範囲で起きるイベントでは，膨大な数のユーザーが同時にツイートをすることでバーストが起こる．このような性質を利用して，リアルタイムで多くのユーザーが注目しているイベントを検出することができる．また，過去のツイートのバーストを解析することで，それぞれのイベントがいつ，どの程度話題になったかを知ることができる．さらに，検出されたイベントを分類することで，より情報の取捨選択が行いやすくなると考えられる．

本稿では，Twitterにおいてユーザーが注目しているイベントを単語単位で検出し，バーストの周期に着目して分類する方法について検討する．ここで，イベントに関連のある語を“話題語”と呼ぶ．本稿では，話題語の抽出にはKleinbergのバースト解析アルゴリズム[4]を用いる．また，イベントには周期的なものと非周期的なものがあると仮定して，得られた話題語を出現周期に着目して分類することを目指す．周期の判別にはピアソン相関係数[5]を用いる．さらに本稿では，その周期に着目することで，イベントがどのように分類されるかを検証する．

2　関連研究

Twitterからのイベント検出は，特に近年活発な研究分野であり，多くの手法が提案されている．Beckerら[6]は，ツイートを内容の類似性に基づいてクラスタリングし，それぞれのクラスタが実世界のイベントに関するものかそうでないかを，教師あり学習で構築した分類器を用いて判別している．渡辺ら[2]は，位置情報付きのツイートを利用して，小規模なローカルイベントの検出を行うシステムの提案をし，さらに実世界の場所と関連の強い場所名を持つツイートに対して位置情報割り当てを行うことで，検出精度を向上させている．Sakakiら[3]は，Twitterのユーザーをソーシャルセンサーと見立て，リアルタイムでのイベントに対するユーザーの反応から，地震や台風の発生と位置を推定する手法を提案している．

また，イベントの周期に着目した研究には，藤木らの手法がある[7]．藤木らは，Kleinbergのバースト解析アルゴリズムを改良した方法で，ブログからリアルタイムで注目語を抽出することを試みている．また，周期的にバーストする語，例えば「正月」などの語が毎回抽出されないように，周期的なバーストを単語の日付特徴などから予測し，抑制している．

3　話題語の抽出と周期の算出

3.1　バースト度に基づく話題語抽出

Kleinbergのバーストアルゴリズムでは，ある期間における各単語のバーストの強さを表す尺度として“バースト度”を用いている[4]．このバースト度が高いほど，その期間において単語 w に関連する話題が注目を浴びているといえる．

Twitterにおける単語 w のバースト度は以下のように求めることができる．解析期間 $t_1, ..., t_n$ において，ツイート集合 $TW_{t_1}, ..., TW_{t_n}$ が送られてくる状況を考える．ここで，TW_{t_k} に含まれるツイートの数を d_{t_k}，そのうち単語 w を含むツイートの数を r_{t_k} とおく．すると，解析期間におけるすべてのツイート数 $D = \sum_{i=1}^{n} d_{t_i}$，単語 w を含むツイート数 $R = \sum_{i=1}^{n} r_{t_i}$ と表すことがで

きる．次に，q_0 を非バースト状態，q_1 をバースト状態として，それぞれの期間に状態 $q_i(i=0,1)$ を与える．非バースト状態 q_0 には，解析期間全体におけるツイート数の期待値 $p_0 = R/D$，バースト状態 q_1 には，p_0 にパラメータ $s(s>1)$ をかけた値 $p_1 = sp_0$ をそれぞれ確率として割り当てる．ツイート集合中における単語 w を含むツイートの出現確率は二項分布に従うと仮定して，期間 t_k において状態 q_i であることに対するコストを，以下の関数により定義する．

$$\sigma(i, r_{t_k}, d_{t_k}) = -\ln\left[\binom{d_{t_k}}{r_{t_k}} p_i^{r_{t_k}} (1-p_i)^{d_{t_k}-r_{t_k}}\right] \quad (1)$$

このとき，期間 t_k における単語 w のバースト度 $bw(t_k, w)$ は以下の式で与えられる．

$$bw(t_k, w) = \sigma(0, r_{t_k}, d_{t_k}) - \sigma(1, r_{t_k}, d_{t_k}) \quad (2)$$

すなわち，バースト度は，状態を q_0 から q_1 としたときのコストの改善度合いによって与えられ，バーストの確からしさを表している．

3.2 ピアソン相関係数を用いた周期判別

ピアソン相関係数を用いた周期判別方法について説明する．ピアソン相関係数は，2 組のデータの間の相関の強さを示す指標である．2 組の数値からなるデータ列 $\{(x_i, y_i)\}(i = 1, 2, ..., m)$ が与えられたとき，ピアソン相関係数は以下のように求められる．

$$C = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \overline{y})^2}} \quad (3)$$

$\overline{x}, \overline{y}$ はそれぞれデータ $\boldsymbol{x} = (x_1, x_2, ..., x_m), \boldsymbol{y} = (y_1, y_2, .., y_m)$ の相加平均である．本稿では，このピアソン相関係数を周期の判別に用いる．

単語 w の出現回数の時系列データ $r_{t_1}, r_{t_2}, ..., r_{t_n}$ が与えられたとき，時間 τ だけ遅らせたときのデータと元のデータとの相関について考える．すなわち，2 組のデータ列 $\{(r_{t_i}, r_{t_i+\tau})\}(i = 1, 2, .., N_w)$ について，以下のようにピアソン相関係数を求める．

$$C(N_w, \tau) = \frac{\sum_{i=1}^{N_w}(r_{t_i} - \overline{r})(r_{t_i+\tau} - \overline{r'})}{\sqrt{\sum_{i=1}^{N_w}(r_{t_i} - \overline{r})^2}\sqrt{\sum_{i=1}^{N_w}(r_{t_i+\tau} - \overline{r'})^2}} \quad (4)$$

ここで，$\overline{r} = \frac{1}{N_w}\sum_{i=1}^{N_w} r_{t_i}, \overline{r'} = \frac{1}{N_w}\sum_{i=1}^{N_w} r_{t_i+\tau}$ であり，N_w は窓幅である．これを用いて，単語 w を含むツイートの出現回数の時間変化が任意の周期 T を持つかどうかを判別するための関数を以下のように定義する．

$$f(T) = \frac{\sum_{i=1}^{N_t} C(T, iT)}{N_t} \quad (5)$$

ただし，$N_t = \frac{n}{N_w}$ である．ここで，窓幅 $N_w = T$ としているのは，周期 T と同じ幅の窓を設け，それを周期倍ずつシフトしたときの相関の度合によって，時間 T ごとに同じ変化の繰り返しが起きているかどうかを判別するためである．周期判別関数 $f(T)$ を，T が小さい順に計算していき，閾値 λ_T を初めて超えたときの T をその単語の周期とし，一度も超えるものがなかった場合は周期なしと判別する．ただし周期は同じような変化の繰り返しが 3 回以上あるものと定義する．したがって，$3T \leqq n$ となる T のみを考慮する．

3.3 部分周期の検出

前節の方法では，閾値を超える $f(T)$ がなければ周期なしと判別される．しかしこの方法では，解析期間の途中から周期性が現れる話題語や，途中から周期性が消える話題語などは検出できない．そこで本節では，このような部分的な周期を判別する方法について述べる．

前節の方法では，窓幅 N_w を解析期間の初めに固定し，もう一方の同じ幅の窓をシフトしながら相関係数を計算しているが，この方法では最初の窓のデータとの相関しか求められず，途中から現れる周期性には対応することができない．そこで，固定する窓を動かすことで，途中から周期性を持つ話題語を判別する．本稿では，3.2 で周期なしと判別された話題語について，周期 T が部分的に存在するかどうかを，以下の手順で判別する．

時系列データ $(r_{t_1}, r_{t_2}, ..., r_{t_n})$ について，$C(T, iT)$ を計算する．$i = 1, 2, 3$ すべてについて $C(T, iT)$ が閾値 λ_T を超えていれば，その単語が部分周期 T を持つと判別する．そうでなければ，固定窓を T だけシフトした状態，すなわち時系列データ $(r_{t_{T+1}}, ..., r_{t_{T+n}})$ に対して上記の判別を繰り返し，$C(T, iT)$ が閾値 λ_T を超えれば，その単語が部分周期 T を持つと判別する．固定窓をシフトしていっても，一度も周期ありと判別されなければ，部分周期 T なしと判別する．これを T が小さい順に計算していき，部分周期 T があると判定された時点で計算を終了する．

4 実験

Streaming API[8] を用いて，2012 年 10 月 5 日〜11 月 5 日の間に収集した日本語のツイートを対象にして，話題語の抽出とその周期に基づく分類を行った．なお，収集した全ツイート数は 15,059,677 であり，抽出の対象とする語は，その語を含むツイートの総数が 50 以上の名詞 29,458 語とした．また，以降では時間の単位を 1 時間とした．図 1 に，この期間における全ツイートの出現回数の変化を示す．本実験では，3.1 で示したバースト度を 1 時間ごとに求め，解析期間全体で最も大きいバースト度をその単語のバースト度とし，バースト度上位 300 語を話題語として抽出した．

この話題語に対し，3.2 で示した方法 ($\lambda_T = 0.5$) で周期の判別を行った結果，$T = 24$(1 日) と判別された単語は 37 語，$T = 168$(1 週間) と判別された単語は 39 語，その他の周期に判別された単語は 6 語であった．表 1 に，$T = 24$ および $T = 168$ と判別された話題語のうち，バースト度の高い上位 10 語を示す．また，その他の周期と判別された話題語を表 2 に示す．

次に 3.3 で示した方法により，上述の例で周期なしと判別された話題語に対して部分周期の判別を行ったところ，30 語が部分周期ありと判別された．結果を表 3 に示す．

5 考察

表 1 を見ると，24 時間周期と判別された話題語は「風呂」,「お昼」,「電車」など，1 日の特定の時間帯に関係のある語が多いことがわかる．ただし,「焼肉」,「お菓子」などの単語は，1 日の特定の時間帯に関係のある語とは言い難い．そこでこれらの語の出現回数の変化を見ると，ある特定の期間のみバーストしていることで，話題語として抽出され，1 日内の出現回数の変化の周期性により，24 時間周期として判別されていることが確認できた．図 1 に示すように，Twitter 全体では，夜の書き込みが多く，朝は少ないという傾向があり,「焼肉」,「お菓子」などの語もこの傾向が表れていたため，24 時間周期と判別されたと考えられる．今後，Twitter 全体のツイート数を考慮して話題語の抽出を行う必要があると考えられる．また，168 時間 (1 週間) 周期と判別された話題語は，ほとんどがアニメやドラマなど，毎週放送されるテレビ番組に関する単語であった.「花火」は，解析期間において土日に多く出現していた．解析期間において，週末に花火大会が行われた地域があったことが推測できる．

一方表 2 から，24 時間周期と 168 時間周期以外の話題語は，ほとんど抽出されていないことがわかる.「今日」,「日本」は，24 時間の周期性が強かったが，12 時間単位の周期性もみられることで，より小さい単位である 12 時間周期と判別された.「適当」,「飯山」は，弱くではあるが 24 時間の周期性がみられたが，閾値によりそれぞれ 96,217 時間の周期と判別された.「リア充度」は，実世界のイベントとは関係がなく，Twitter 上で「リア充度」に関係するツイートが偶然周期的にバーストしたものと考えられる．

最後に，表 3 より，部分周期があると判定された話題語について,「クリスマス」,「ハロウィン」のようにあらかじめ起こることのわかっているイベントや,「サッカー」,「野球」のようにある特定の期間に定期的に行われるイベントが抽出されていた．また,「銀さん」,「がれ」は，テレビ番組に関する単語であるが，解析期間の途中から現れたり，通常とは異なる時間帯にバーストが起きたために，部分周期として判別された．23 時間周期と判別されている話題語がいくつかみられるが，24 時間周期のものが 23 時間周期と判別されたと思われる．周期性の判別方法について，閾値の設定や手法の改善を行う必要があると考えられる．

6 おわりに

本稿ではバースト度に基づいて話題語を抽出し，ピアソン相関係数を用いてその周期を判別する方法について検討した．周期に基づく分類結果から，抽出された周期のほとんどは 24 時間周期と 1 週間周期の語であり，それぞれに特徴的な話題語が存在することを示した．また，部分的な周期を判別する方法について述べ，その判別結果を示した．

今後は，周期の判別精度をより向上させると共に，単語の共起関係や，出現回数の時系列変化の類似性などに着目することで，話題語を集約する方法について検討していく．また，部分周期については，部分周期が発生している期間に着目した解析についても行っていく予定である．

参考文献

[1] Twitter: http://twitter.com/
[2] 渡辺一史, 大知正直, 岡 部誠, 相知理紀夫, Twitter を用いた実世界のローカルイベント, 第 4 回楽天研究開発シンポジウム, 2011.
[3] T. Sakaki, M. Okazaki, and Y. Matsuo: Earthquake shakes twitter users: Real-time event detection by social sensors, Proceedings of 19th

International Conference on World Wide Web, pp.851—860, 2010
[4] J. Kleinberg: Bursty and Hierarchical Structure in Streams, Proceeding of the 8th ACM SIGKDD In- ternational Conference on Knowledge Discovery and Data Mining, 2002
[5] 金明哲:テキストデータの統計科学入門, 岩波書店,244pp,2009
[6] Hila Becker, Mor Naaman, Luis Gravano: Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM 2011
[7] 藤木稔明, 奥村学:周期的に発生する burst の予測と抑制, 人工知能学会, 第 73 回知識ベースシステム研究会, 2006
[8] GET statuses/sample: https://stream.twitter.com/1/statuses/sample.json

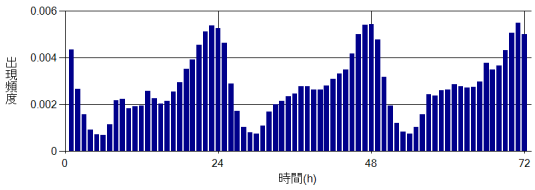

図 1 : 全ツイートの出現回数の変化

表 1: $T = 24, 168$ と判別された話題語の一部

T=24	T=168
明日	マギ
バイト	#precure
風呂	#chu2koi
お昼	#hidamari
電車	コナン
遅刻	銀魂
夕焼け	ジョジョ
お菓子	#sao_anime
焼肉	花火
映画	月曜日

表 2: $T = 24, 168$ 以外の周期と判別された話題語

周期 T	話題語
12	今日
12	日本
96	適当
217	飯山
239	リア充度

表 3: 部分周期 T を持つと判別された話題語

部分周期 T	話題語
12	サッカー
12	日本一
12	紹介
12	中日
12	一部
12	自慢
12	いたずら
12	シュート
15	地震
23	ハロウィン
23	優勝
23	予測変換
23	おでん
23	クリスマス
23	簡単
23	仮装
23	相性
23	もも
24	巨人
24	たけ
24	野球
24	常識
24	ポテト
24	文字
24	イラスト
24	食べ物
24	チョコ
24	爆発
168	銀さん
168	がれ