2E1-1

Twitterにおけるキーワードの出現周期の自動判別手法に関する検討

A Study on Periodicity Discrimination Method for Keyword Frequency in Twitter

佐々木謙太朗 *1 Kentaro Sasaki
吉川大弘 *2 Tomohiro Yoshikawa
古橋武 *2 Takeshi Furuhashi

名古屋大学工学研究科
Graduate School of Engineering Nagoya University

Twitter is rapidly prevailing on the Web society in recent years. A lot of Twitter contents contain periodicity, and it is meaningful to find the periodicity from them. In this paper, we propose the method to extract the periodicity and estimate its range based on correlation coefficient from queries or keywords in Twitter contents. We apply the proposed method to the keywords representing hot topics in Twitter and show the validity of it.

1. はじめに

近年，Twitter[*1] に代表されるマイクロブログサービスや，ソーシャルネットワーキングサービス (SNS) が広く普及し，多くのユーザーが Web 上に情報を発信するようになった．またその他にも，多くのユーザーが，Web 検索エンジンや，コミュニティQA(CQA) を用いることによって，日常生活における疑問などを解決するための情報を入手している．Web 検索エンジンや CQA においては，ユーザーが使用するクエリや質問記事を作成することで，これらのサービスを通じてユーザーの情報要求が発信されているととらえることもできる．ユーザーの発信する情報は常に変化し続けるが，それらの変化には周期性が存在するものも多い．検索エンジンに入力されるクエリの周期性を対象とした研究は特に盛んに行われている [Vlachos 2004][Shokouhi 2012]．例えば,「クリスマス」というクエリは毎年 12 月に検索頻度が増加するが，それ以外の期間ではあまり検索されないといった特徴がある．このような検索の周期性を特定することによって，時期に応じたクエリ拡張が可能となる．同様の研究は CQA においても行われており，大塚ら [大塚 2013] は時系列トピックモデルとフーリエ変換を用いて周期的に変化するトピックの特定を行っている．

マイクロブログサービスや SNS においても,「クリスマス」のような特定のキーワードの出現頻度は周期的に変化する．Fukuhara ら [Fukuhara 2005] は，ブログにおける人々の関心パターンを 5 つに分類しているが，そのうちの 1 つとして周期型を挙げている．藤木ら [藤木 2006] は，ブログからリアルタイムで出現頻度の急増する注目語を抽出することを試みている．また [藤木 2006] においては，周期的に出現頻度の増加する語は，抽出されても情報量が少ないとして，周期的な出現頻度の急増を予測し，抑制している．

以上のように，Web コンテンツから周期的な性質を特定する試みがこれまで数多く報告されている．しかし，既存の研究の多くは，想定する周期を 1 週間単位，1 年単位などに限定していたり，周期性が対象とする期間全体で現れることを前提としている．そこで，本稿では，Web コンテンツにおけるクエリやキーワードの出現頻度から，任意の周期の周期性を特定し，かつ周期性の現れる範囲も考慮した手法として，ピアソンの相関係数 [金明哲 2009] を用いた手法を提案する．提案手法を Twitter におけるキーワードに対して適用し，その有効性を示す．

Twitter は 140 文字以内という制限により，ユーザーが気軽にメッセージを投稿でき，リアルタイム性が高い．そのため，ユーザーの興味を惹く何らかのイベントが発生すると，それと同期して関連するツイートが急激に増加するという傾向がある．このような現象をバーストと呼ぶ [Kleinberg 2002]．例えば，地震が発生した際には,「地震」という単語を含むツイートがバーストする．本稿では，周期性を持つバーストの抽出と，その範囲の特定を試みる．ここで，対象とする期間においてバーストしている単語を “話題語” と呼ぶことにする．提案手法において，話題語は，Kleinberg のバースト解析アルゴリズム [Kleinberg 2002] を用いて抽出する．そして，抽出された話題語に対して周期性の判別を行う．また本稿では，周期性に着目することにより，話題語の特性を把握できることを示す．

連絡先: 連絡先:佐々木謙太朗，名古屋大学工学部工学研究科，名古屋市千種区不老町，052-789-2793，052-789-3166，sasaki@cmplx.cse.nagoya-u.ac.jp

*1 http://twitter.com/

2. Kleinberg のバースト解析アルゴリズム

本節では，Kleinberg の手法について説明する．Kleinberg の手法を用いることによって，離散時間で文書 (ツイート) が送られてくる状況において，ある単語を含む文書がバーストしている期間（バースト期間）と，バーストしていない期間（非バースト期間）を自動で判別することができる．また，ある期間における単語のバーストの強さを “バースト度” として定量化することができる．

2.1 バースト期間・非バースト期間の決定

Twitter における，単語 w を含むツイートのバースト期間，非バースト期間は以下のように決定される．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 $D = \sum_{k=1}^{n} d_{t_k}$，単語 w を含むツイート数 $R = \sum_{k=1}^{n} r_{t_k}$ と表すことができる．次に，期間 t_k に対して，q_{0_k} を非バースト状態，q_{1_k} をバースト状態として，解析期間 $t_1, ..., t_n$ に対する状態系列 $\mathbf{q} = (q_{i_{t_1 1}}, q_{i_{t_2 2}}, ..., q_{i_{t_n n}})(i_{t_k} = 0, 1)$ を与える．非バースト状態 q_{0_k} には，解析期間全体における単語 w を含むツイートの出現の期待値 $p_0 = R/D$，バースト状態 q_{1_k} には，p_0 にパラメータ s をかけた値 $p_1 = sp_0$ を

それぞれ確率として割り当てる．ただし，s は $s > 1$ であり，$p_1 \leqq 1$ を満たすものとする．ツイート集合中における単語 w を含むツイートの出現数は二項分布に従うと仮定して，期間 t_k において状態 $q_{i_{t_k}k}$ 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．

$$\sigma(i_{t_k}, r_{t_k}, d_{t_k}) = -\ln\Big[\binom{d_{t_k}}{r_{t_k}} p_{i_{t_k}}^{r_{t_k}} (1-p_{i_{t_k}})^{d_{t_k}-r_{t_k}}\Big] \quad (1)$$

このコストが最小になるように状態系列 $\mathbf{q}$ を決定することで，各期間の状態に応じて非バースト期間，バースト期間を定めることができる．しかし，単語 w を含むツイートの数が閾値付近で変化すると，不自然な状態の切り替えが生じてしまうため，これに加えて，状態 $q_{i_{t_k}k}$ から状態 $q_{i_{t_{(k+1)}}(k+1)}$ への状態の遷移を妨げる関数を以下のように定義する．

$$\tau(i_{t_k}, i_{t_{(k+1)}}) = \begin{cases} (i_{t_{(k+1)}} - i_{t_k})\gamma \ln n & (i_{t_{(k+1)}} > i_{t_k}) \\ 0 & (i_{t_{(k+1)}} \leqq i_{t_k}) \end{cases} \quad (2)$$

ここで，γ は $\gamma > 0$ を満たすパラメータである．これらを合わせて，状態系列 $\mathbf{q}$ に対するコスト関数は以下の式で与えられる．

$$c(\mathbf{q}|r_{t_k}, d_{t_k}) = \Big(\sum_{k=1}^{n-1} \tau(i_{t_k}, i_{t_{(k+1)}})\Big) + \Big(\sum_{k=1}^{n} \sigma(i_{t_k}, r_{t_k}, d_{t_k})\Big) \quad (3)$$

このコスト関数を最小にする状態系列 $\mathbf{q}$ を決定することで，非バースト期間，バースト期間を定める．

なお本稿では，上述のパラメータ $s = 2$，$\gamma = 1$ とする．

2.2 単語のバースト度

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

$$bw(t_k) = \sigma(0, r_{t_k}, d_{t_k}) - \sigma(1, r_{t_k}, d_{t_k}) \quad (4)$$

すなわち，バースト度は，状態を q_{0k} から q_{1k} としたときのコストの改善度合いによって与えられ，その期間におけるバーストの強さを表している．

3. 提案手法

本節では，キーワードやクエリなどの Web コンテンツの時系列データから，周期性を自動判別する手法について述べる．なお本稿では，「周期性がある」とは，特定の時間間隔で，バースト度に対して類似した変化が 3 回以上繰り返されることと定義し，このときの時間間隔を “周期” と呼ぶことにする．

3.1 ピアソンの相関係数

ピアソンの相関係数は，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．

$$C = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{m}(y_i - \overline{y})^2}} \quad (5)$$

$\overline{x}, \overline{y}$ 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，このピアソンの相関係数を時系列データへ拡張し，周期性の抽出に用いる．

時系列データ $z_{t_1}, z_{t_2}, ..., z_{t_n}$ が与えられたとき，窓幅 N_w として，解析期間の始めから時間 a だけシフトした位置の固定窓と，同じ窓幅で固定窓の位置から時間 b だけシフトした比較窓との相関の強さを求める．すなわち，2 組のデータ列 $\{(z_{t_{i+a}}, z_{t_{i+a+b}})\}(i = 1, 2, .., N_w)$ について，以下のようにピアソンの相関係数を求める．

$$C(N_w, a, b) = \frac{\sum_{i=1}^{N_w}(z_{t_{i+a}} - \overline{z})(z_{t_{i+a+b}} - \overline{z'})}{\sqrt{\sum_{i=1}^{N_w}(z_{t_{i+a}} - \overline{z})^2}\sqrt{\sum_{i=1}^{N_w}(z_{t_{i+a+b}} - \overline{z'})^2}} \quad (6)$$

ここで，$\overline{z}, \overline{z'}$ はそれぞれ固定窓，比較窓におけるデータの相加平均である．

3.2 周期判別得点 $Sc(T)$ の算出

時系列データに拡張したピアソンの相関係数 ($C(N_w, a, b)$) を用いて，周期 T に対する周期性の強さを定量的に評価する方法について説明する．

本稿では，窓幅 $N_w = T$ として，比較窓を周期倍ずつシフトし，相関係数の値が何回連続で閾値を超えたかによって周期 T での周期性の強さを与える．今，(6) 式における $C(T, j_1T, j_2T)$ を考える．以下に，周期 T での周期性の強さを表す周期判別得点 $Sc(T)$ を求める手順を示す．

1) $j_1 = 0$ として，固定窓を解析期間の始めに設置する．

2) $j_2 = j_1 + 1$ から j_2 を小さい順に変化させ，比較窓を周期 T 倍ずつシフトしながら，$C(T, j_1T, j_2T)$ の値を計算する．

3) $C(T, j_1T, j_2T)$ の値が，閾値 λ_T を連続して超えた回数の最大値を n_{j_1} とする．ただし，$j_2 = 1$ において閾値を超えた場合は 1 回余分に数える．

4) $j_1 = j_1 + 1$ として固定窓を一周期分シフトし，2)，3) の手順を繰り返す．

5) すべての j_1 について n_{j_1} を求め，n_{j_1} の最大値を n_{max} とする．

6) 周期判別得点 $Sc(T)$ を，以下の式により求める．

$$Sc(T) = \begin{cases} \frac{n_{max}}{N_T} & (n_{max} \geqq 3) \\ 0 & (n_{max} < 3) \end{cases} \quad (7)$$

ただし，$N_T = \frac{t_n}{T}$(解析期間における窓幅 T としたときの窓の数) である．分母の N_T は，$Sc(T)$ の値が 0 から 1 の間の値をとるようにするための正規化項である．周期判別得点 $Sc(T)$ は，値が大きいほど，その周期の周期性が現れる範囲が広いことを意味し，1 ならば解析期間全体に周期 T での周期性があり，0 ならば全く周期性が現れていないことを示す．この周期判別得点 $Sc(T)$ を，解析期間において想定されるすべての周期について計算することで，単語 w に対するバーストの周期性を抽出する．なお，$n_{max} < 3$ のときに $Sc(T) = 0$ としているのは，周期性を，類似したバースト変化の 3 回以上の繰り返しと定義しているためである．また，同じ理由から，考慮する周期は $3T \leqq t_n$ を満たす T とする．

4. 実験

話題語のバースト度の変化に対して提案手法を適用し，得られた周期判別得点に基づいて話題語のクラスタリングを行った．実験には，Streaming API[*2] によって，2012 年 10 月 5 日〜11 月 15 日の間に収集した日本語のツイートを用いた．なお，収集した全ツイート数は 19,791,138 であり，抽出の対象とする単語は，その単語を含むツイートの総数が 100 以上の名詞 10,900 語とした．また，以降では 1 期間の単位を 1 時間とした．

4.1 実験手順

4.1.1 話題語の抽出

2.2 で述べたように，バースト度が高い単語は，その期間において注目を浴びている単語であるといえる．そこでここでは，上述の対象語句のうち，バースト度の高い順に 300 語を話題語として抽出し，その周期を求めた．以下に抽出の手順を示す．

1. 単語 w のそれぞれの期間（時間）におけるバースト度 $(bw(t_k))$ を算出する．
2. バースト度の最大値を単語 w のバーストスコア $bw_{max}(w)$ とする．
3. $bw_{max}(w)$ の値が高い上位 300 語を話題語として抽出する．

4.1.2 ノイズ除去

ここでは，ツイートの出現頻度そのものよりも，バーストの強さを表すバースト度を用いた方が，より適切にバーストの周期性を抽出できると考えられるため，バースト度の時間変化を用いて周期性を抽出した．さらに，以下の方法により，バーストしていない期間をノイズとして除去した．2.1 で述べた手法により，単語 w の出現頻度の時間変化に対して，バースト期間と非バースト期間を決定し，非バースト期間のバースト度を 0 として計算し，バースト度の時系列データ $bw(t_1), bw(t_2), ..., bw(t_n)$ を得た．

4.1.3 周期性の抽出

話題語 300 語におけるバースト度の時系列データ $bw(t_1), bw(t_2), ..., bw(t_n)$ に対し，3.2 で示した方法 $(\lambda_T = 0.8)$ で，周期判別得点 $Sc(T)$ を算出した．ただし，24 時間未満の周期で周期性が現れることは考えにくいため，$T \geqq 24$ を満たす周期についてのみ計算を行った．

4.1.4 クラスタリング

周期判別得点を素性として，話題語のクラスタリングを行った．クラスタリング手法は k-means 法を用いた．距離尺度としてはユークリッド距離を用い，クラスタ数は 10 とした．

4.2 結果と考察

クラスタリングの結果を表 1 に示す．なお，空のクラスタ（クラスタに属するデータ数=0 ）は除外している．以下にそれぞれのクラスタについての考察を述べる．

4.2.1 クラスタ 1,6

クラスタ 1 に属する「最初」,「出会い」という話題語については，$Sc(25) = 0.08$ で，25 時間周期以外の周期判別得点は 0 であった．また，クラスタ 6 に属する「一文字」については $Sc(72) = 0.21$ であり，それ以外については 0 であった．これは，それぞれの話題語において，25 時間周期，72 時間周期で一部期間バーストしていたことを示している．「一文字」のバースト度の時間変化を図 1 に示す．これら「最初」,「出会い」,「一文字」の話題語はすべて，それが含まれるハッシュタグ（そのツイートが何の話題に関することかを明示的に示すもので,「#○○」といった形でツイートの文末に付与される）の流行によるものであった．例えば,「#名作のタイトルに一文字足すとよく分からなくなる」というハッシュタグとともにあるユーザーが発言すると，それに触発されて他のユーザーも発言をするといった形で，ハッシュタグそのものがバーストする．これは基本的に実世界のイベントとは関係がなく，周期性が見られたのは偶然によるものと考えられる．

4.2.2 クラスタ 2

クラスタ 2 に属する話題語は，図 2 のように解析期間全体に渡って 1 日周期の周期性をもつ話題語であった．これらの話題語は,「お昼」や,「おやつ」,「バイト」などのように，日々の生活における特定の時間帯でのユーザーの興味・関心に関連した語句であると考えられる．また,「恋愛運」などの話題語は，その日の運勢を占うツイートに含まれるものであるため，毎日早朝にバーストし，1 日周期の周期性として抽出された．

4.2.3 クラスタ 3

クラスタ 3 に属する話題語は，図 3 に示すような周期性をもち，周期判別得点は $Sc(168)$ が 1 に近い値，168 時間 (=1 週間) 未満の周期についてはそれよりやや小さい値であった．これは，1 週間周期が解析期間全体に現れており，その他の周期（図より，1 日周期）が一部の期間に現れていることを示している．これらの話題語は,「電車」や「遅刻」など，学校や仕事と強く関連した語であり，平日と休日とでバーストの仕方が異なる，1 週間の生活サイクルに関連している話題語であると考えられる．

4.2.4 クラスタ 5

クラスタ 5 に属する話題語は，図 4 のように，解析期間における一部の期間でのみ 1 日周期の周期性が現れる傾向があった.「巨人」,「野球」,「中日」などといった話題語は，プロ野球のクライマックスシリーズに関連して出現したと思われる単語であり，解析期間において同じような時間帯に連続して野球の試合が行われたことで，周期性が現れたと考えられる．

4.2.5 クラスタ 7

クラスタ 7 に属する話題語は，図 5 のような 1 週間周期の周期性をもち，ほとんどの単語において $Sc(168)$ 以外の周期 T に対する周期判別得点 $Sc(T)$ が 0 であった．これらの話題語はすべて，毎週放送されるテレビ番組に関連していた．ただし,「明日」と「学校」については，図 6 に示すように，$T = 168$ 以外の周期についても周期判別得点が 0 でなく，クラスタ 3 のような 1 週間の生活サイクルに起因すると考えられるバーストの傾向が表れていた．クラスタ 3 ではなくクラスタ 7 に属した理由としては，クラスタ 3 の話題語よりも，1 週間周期の周期性がより強く現れていたためであると考えられる．

5. おわりに

本稿では，Web コンテンツにおけるクエリやキーワードの出現頻度から，任意の周期性を自動判別する手法を提案した．Twitter における話題語に対して提案手法を適用し，周期判別得点を素性としたクラスタリングを行うことによって，周期の長さや周期性の現れる範囲に基づくクラスタが生成されることを確認した．さらに提案手法により，期間全体に現れる周期性だけでなく，一部期間でのみ現れる周期性をもつ話題語も抽出できることを示した．また，周期性に着目することにより，

*2 https://dev.twitter.com/docs/api/1.1/get/statuses/sample

表 1: クラスタリング結果

クラスタ 1	最初，出会い
クラスタ 2	お昼，バイト，おやつ，定期，わたし，今日，恋愛運，金運，運勢，健康運
クラスタ 3	電車，定期，遅刻，昨日
クラスタ 4	誕生日，ラッキーアイテム
クラスタ 5	巨人，野球，中日，簡単，ジャイアンツ，ドラマ
クラスタ 6	一文字
クラスタ 7	相棒，マギ，グルメ，大奥，アラジン，明日，#chu2koi，学校，アリババ，銀魂，モルさん，#magi，がれ，悪夢ちゃん，孤独，アメトーク，ジョジョ，#litbus_anime，プリキュア，合体，結婚，アリババくん，#sao_anime，高校入試，金魂，ヨシヒコ，れい，嵐ちゃん，勇者ヨシヒコ，#jojo_anime，ひだまり，#oniai
クラスタ 8	上記以外の話題語

各話題語がどのような話題に起因してバーストしているのかといった特性の把握を支援できることを示唆した．

今後は，Twitter だけでなく，検索クエリログやその他のブログ等にも提案手法を適用し，提案手法について検討していく予定である．また，フーリエ変換など，従来手法との比較も行っていく予定である．

参考文献

[Vlachos 2004] Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. Identifying similarities, periodicities and bursts for online search queries. Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data(SIGMOD'04), pp. 131-142 (2004)

[Shokouhi 2012] Milad Shokouhi and Kira Radinsky. Time-sensitive query auto-completion. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'12), pp. 601-610 (2012)

[大塚 2013] 大塚淳史, 関洋平, 佐藤哲司:話題の周期性に着目した情報要求言語化のためのクエリ拡張手法の提案, 第 5 回データ工学と情報マネジメントに関するフォーラム DEIM2013 論文集, C9-1 (2013)

[Kleinberg 2002] J. Kleinberg: Bursty and Hierarchical Structure in Streams, Proceeding of the 8th ACM SIGKDD In- ternational Conference on Knowledge Discovery and Data Mining (2002)

[金明哲 2009] 金明哲:テキストデータの統計科学入門, 岩波書店,244pp (2009)

[Fukuhara 2005] T. Fukuhara, T. Murayama, and T. Nishida: Analyzing concerns of people using Weblog articles and real world temporal data, in WWW 2005 Chiba, Japan,pp. 1-12 (2005)

[藤木 2006] 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,Vol26 (2006)

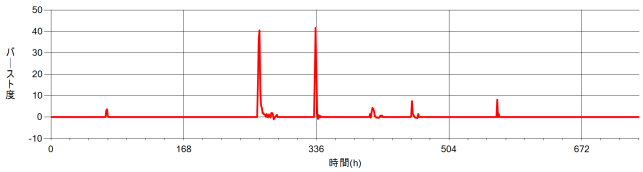

図 1: 「一文字」のバースト度の時間変化

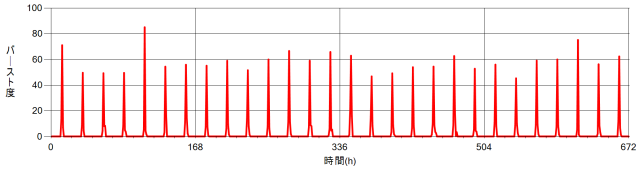

図 2: 「お昼」のバースト度の時間変化

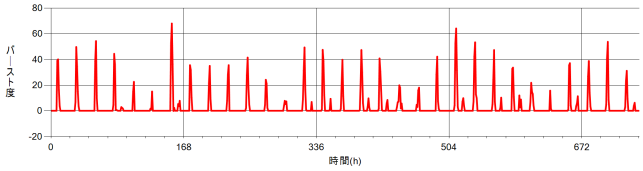

図 3: 「電車」のバースト度の時間変化

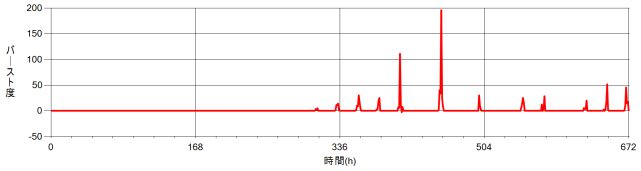

図 4: 「巨人」のバースト度の時間変化

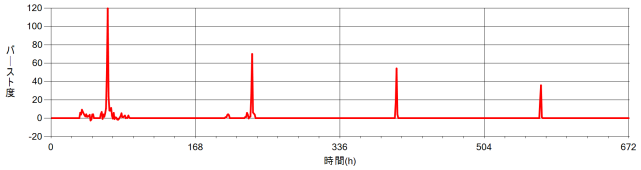

図 5: 「マギ」のバースト度の時間変化

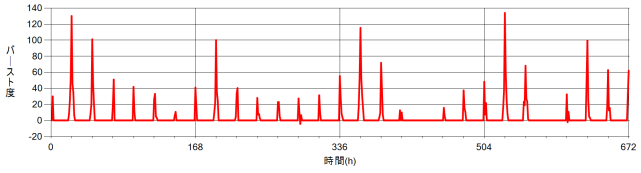

図 6: 「明日」のバースト度の時間変化