2H4-OS-15a-3

ツイッターストリームにおけるバースト時間スケールの同定

Identifying characteristic time-scales of the burst in Twitter stream

橋本康弘[*1]
HASHIMOTO, Yasuhiro

[*1]東京大学新領域創成科学研究科人間環境学専攻
Department of Human and Engineered Environment Studies, Graduate School of Frontier Sciences, The University of Tokyo

Bursty behaviors are frequently observed in the web communication such as social networking services and contents sharing services. They show an abrupt surge of the number of message/query posting within a certain time period, while some of them are induced by external stimuli such as TV news, seasonal cultural activities and natural phenomena, and others are endogeneously formed by inter-personal communications such as rumor-spreading. Each bursting phenomenon has a unique temporal pattern according to its intrinsic growth dynamics. It is important to identify their characteristic time-scales for the understanding of human and web interaction. We discuss the method on how to tackle that problem.

1. はじめに

現在，ウェブ上に広がったソーシャルネットワークにおいて日々膨大なデータが生成され流通している．それはあたかも人間という活動体を介して実世界のさまざまな出来事，例えば自然現象やあるいは人間社会の事象が知覚され，データ化され，それがウェブと人間によって構成される一つの巨大な系に刺激として取り込まれているように見える．生物の脳の機能が神経発火のパルスの内部的な流通によって発揮されるのと同様，ウェブを流通する情報によって何らかの全体的な機能が発揮され得るのかという素朴な疑問が生じる．実際，SNS 上で生じるメッセージ生成は時折集団的な振る舞いを示し，それらは「バースト」と呼ばれ実用や学術的な観点から研究の対象となっている．外部からの明示的な入力がなくてもある種の特徴的なパルスを生成する脳システムのように，実世界からの刺激によって内部状態を遷移し続ける巨大システムとしてウェブ-人間の結合系を理解できるのではないか．そのための一つの手がかりとして我々は SNS の出力が示すバースト現象に着目する．本研究ではツイッターから得られる時系列の頻度データを分析し，多様な時間構造を内包するバースト現象から特徴的な時間スケールを同定する方法について検討する．

2. 関連研究

Crane らはバーストのメカニズムが小さな外乱に対して内的に自己励起されるものと，大きな外乱に対して即応的に立ち上がるものに分けて議論した [Crane 08]．また，Lheman らはツイッターデータで観測されるバーストの時間的特徴を用いてトピックのカテゴライズを試みた [Lehman 12]．時系列からのバースト検知そのものも重要な研究テーマであり，Kleinberg らはオートマトンの状態遷移モデルとしてバースト生成を記述した [Kleinberg 02]．時系列からバーストを切り出すためにウェーブレット解析的な手法 [Zhu 03, Zhang 06] や，異なる時間スケールの移動平均の差分を利用した手法 [Vlachos 04, He 10] なども提案されている．我々はバースト現象をトレンド分析のような実用面ではなく，ウェブと人間の集合的振る舞いを理解する手がかりとして活用したい．そのためにはバーストを時系列から合理的に切り出すための考え方について慎重に検討する必要がある．

連絡先: 橋本康弘，hashimoto@vis.k.u-tokyo.ac.jp

3. 分析

3.1 対象データ

2011 年 7 月から 2013 年 3 月までの約 1 年半分のツイッターデータ（サンプリング）を用いた．ツイッターストリームには様々なトピックが混在し，それぞれが異なる時間構造を持つ．例えば「おはよう」「おやすみ」といった単語が日周期を持つ一方で,「地震」のような突発的な事象に関連した単語では数秒から数分の短いスケールで急激なツイート数の増加が見られる．後者のような事象について大きな時間単位で時系列を積算すると，固有の特徴を見落とす可能性がある．

3.2 時系列の作成

ツイッターデータはツイートと呼ばれる時刻情報を持った短文のストリームであり，まずそこから単位時間あたりのツイート量の時系列を構成する．時系列の構成はガウシアンカーネルを用いた一般的な密度推定の方法を用いる．

$$x(t) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i:|T_i - t| < 3\sigma} \exp\left[\frac{-(T_i - t)^2}{\sigma^2}\right] \qquad (1)$$

t はある時間幅 Δt で離散化された時刻，$x(t)$ は時刻 t におけるツイート密度，T_i は i 番目のツイートが生成された時刻である．平滑化パラメータ σ はツイートに対して仮定される時間的広がりであり，ここでは $\sigma = 3\Delta t$ と固定した．これが実質的に時系列の解像度＝時間スケールを決定する．様々な事象に対して時系列を構成するために，ここでは単純化して特定の単語を含むツイートを抽出して上記の方法を適用する．

3.3 時間スケールの同定

各単語に固有の時間スケールを推定するための考え方は単純である．今，ある単語について「固有の時間スケールを反映した真の時系列」というものを仮定したとき，その時間スケールよりも大きなスケールで眺めた場合には固有の特徴は失われるはずである．逆に小さなスケールで眺めた場合には真の時系列の周りで揺らぐ時系列として観測されるはずである．したがって，着目する時間スケール Δt_S で平滑化された時系列と，

それよりも長い時間スケール Δt_L で平滑化された時系列の差分を調べることで，単語固有の時間スケールに関する知見が得られるのではないか，というのが基本的なアイデアである．

図 1 にいくつかの単語について，異なる時間スケールの差分について得られた確率分布を示す．横軸はスケールされた差分

$$y(t) = \frac{x_S(t) - x_L(t)}{\sqrt{x_L(t)}}, \tag{2}$$

を表す．case1 から case6 はそれぞれ異なる 2 つの時間スケール：(1)10sec/1min，(2)1min/10min，(3)10min/1hour，(4)1hour/4hour，(5)4hour/1day，(6)1day/3day の差分に対応する．時間解像度を $\sigma = 3\Delta t$ が決定することを踏まえると，例えば (4) は小スケールが 6 時間，大スケールが 24 時間に相当し，日周期をまたいだ時系列の比較となる．規則的で強い日周期を持つ “おやすみ”，不規則で緩やかな変動を示す “雨”，不規則で急激な変動を示す “地震” という 3 つの単語について，図 1 はそれぞれが異なるスケール間の振る舞いを示すことを明らかにしている．例えば “おやすみ” の例では (3) の 1 時間と 6 時間の間，および (4) の 6 時間と 24 時間の間でスケール間の相対的なズレが大きくなり，この辺りに特徴スケールが存在することを示唆している．一方，“地震” の例ではスケールを変えても分布の形はあまり変化せず，特徴的な時間スケールが明確には存在しないことを示唆している．

4. まとめ

本稿では一部の単語についてスケール間の差分の確率分布を求め，関連する事象について特徴的な時間スケールが異なることを示した．定性的には違いが見えたが，“地震” の例のように事象によっては具体的に単一の固有スケールが与えられるとは限らないことも示された．以上の知見に基づいて会場ではバースト検知の考え方について議論したい．

参考文献

[Crane 08] Crane, R., Sornette, D.: Robust dynamic classes revealed by measuring the response function of a social system, *PNAS*, 105, 41, pp. 15649-—15653, (2008).

[Lehman 12] Lehmann, J., et al.: Dynamical classes of collective attention in twitter, *Proc. of the 21st Int. Conf. on World Wide Web*, pp. 251–260, (2012).

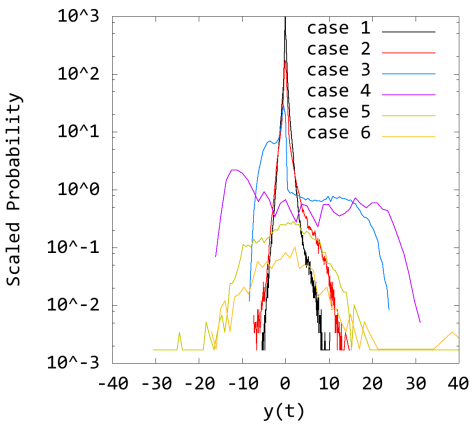
[Kleinberg 02] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 91–101, (2002).

[Zhu 03] Zhu, Y., Shasha, D.: Efficient elastic burst detection in data streams, *Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 336–345, (2003).

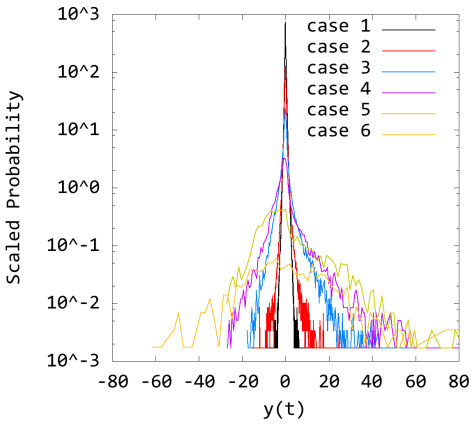
[Zhang 06] Zhang, X., Shasha, D.: Better Burst Detection, Proc. of the 22nd Int. Conf. on Data Engineering, pp. 146, (2006).

[Vlachos 04] Vlachos, M., et al.: Identifying similarities, periodicities and bursts for online search queries *Proc. of the 2004 ACM SIGMOD Int. Conf. on Management of Data*, pp. 131–142, (2004).

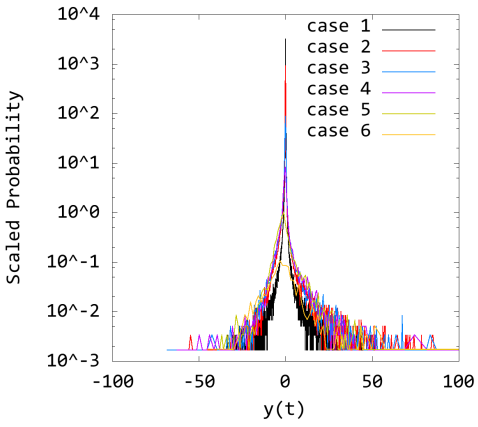
[He 10] He, D., Parker, D. S.: Topic dynamics: an alternative model of bursts in streams of topics, *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 443–452, (2010).



“おやすみ”



“雨”



“地震”

図 1: 異なる時間スケールで構成した 2 つの時系列間の差分の確率分布．