4I2-5

Hawkes Processと最適区間幅推定を用いたWeb Dataの解析

Web data analysis with hawkes process and histogram bin-width optimization

三宅 雅矩 *1　池上 高志 *1　岡 瑞起 *2　橋本 康弘 *2
Masanori Miyake　Takashi Ikegami　Mizuki Oka　Yasuhiro Hashimoto

*1東京大学大学院総合文化研究科
Graduate School of Arts and Science, The University of Tokyo

*2筑波大学システム情報系
Department of Computer Science, University of Tsukuba

Many studies on statistical properties of hawkes process, which is the self-excitable extension of poisson point process, have been conducted these days, and there are a lot of applications to timeseries-analyses in fields such as seismology, neuroscience, financial trading, and web science. There are also theoretical studies about the relationship between the excitability of hawkes process and its burstiness, and the existence of burst transition has been proven with simulations. But there has been no demonstration of the burst transition using real data. We conducted timeseries-analyses of social web data with hawkes process, and we found there is a critical point concerning its excitability, and burst transitions occur around the point.

1. はじめに

自己励起型の点過程である Hawkes Process を用いた地震や神経細胞の活動，金融取引パターンなどの解析および予測が従来より盛んに行われてきた [Filimonov 12]．近年では非常に多数の要素が互いに影響を及ぼし合うような系の内部のダイナミクスを上手く捉えられるという特性を活かし，Social Networking Service(SNS) における人々の活動を Hawkes Process によってモデル化するなどといった研究も行われている [Oka 15]．

一方で，単純な Poisson 点過程とは全く異なる挙動を示すこの Hawkes Process の性質について，理論的に解明しようとする試みもなされてきた．特に，Hawkes Process におけるイベントの励起率について，ある値を境に Burst の生じない静的な定常状態から，Burst が頻繁に生じる非定常状態へと移行 (burst transition) するような critical point が存在することが理論的に示され， シミュレーションによる実験でもそれに従った結果が得られている [Onaga 14]．

本研究では，SNS から得られた実データについて Hawkes Process を用いた fitting を行うとともに，データ時系列からヒストグラムを作成する上での最適区間幅推定を通して非定常状態の程度を測定した．これらの結果より，実データにおいても Hawkes Process の励起率に関して critical point が存在し，その点の周辺を境に burst transition が生じることを実証することができた．

2. 解析手法と対象

2.1 Hawkes Process

Hawkes Process は正のフィードバックを考慮して Poisson Process を拡張した自己励起型の点過程である [Hawkes 71]．1 次元 Hawkes Process において，イベントの発生率 (強度関数)$\lambda(t)$ は過去に生じたイベントとの時間的関係によって変化し，以下の式で表される．

$$\lambda(t) = \rho + \alpha \sum_{t_k<t} f(t - t_k) \tag{1}$$

ここで，ρ は常に保たれているベースラインの発生率であり，α はイベント発生率のジャンプ幅を示す．

連絡先: 三宅雅矩，miyake@sacral.c.u-tokyo.ac.jp

指数関数カーネル $f(x) = e^{-\beta x}$ を用いると，

$$\lambda(t) = \rho + \alpha \sum_{t_k<t} e^{-\beta(t-t_k)} \tag{2}$$

のように表現され，この形の Hawkes Process は最尤推定における利点などから多くの実データの解析において用いられている．

上記の指数関数カーネルを用いた Hawkes Process の内部でのダイナミクスを定量化する指標として，以下に示される branching ratio が頻繁に用いられる．

$$\text{branching ratio} = \int_0^\infty \alpha e^{-\beta t} dt = \frac{\alpha}{\beta} \tag{3}$$

この branching ratio はある一つのイベントから内的に生成される別のイベントの平均個数に対応し，この値が 1 より大きくなると非定常状態になることは，従来より様々なデータを用いて確認されている．

2.2 最適区間幅推定

神経細胞のスパイク時系列などといったデータに関して，そのイベントの生成頻度 (レート) を推定するため，適当な時間幅をもつ区間に分割した時系列内でのイベント生成率からヒストグラム (Peristimulus Time Histogram, PSTH) を作成することがよくある．この時間幅に関しては，平均二乗誤差最小化の観点から最適区間幅を決定する手法が考案されており，以下の手順に従って求められる [Shimazaki 07]．

1. n 回の試行より得られたイベント時系列について，その観測期間 T を幅 Δ の N 個の bin に区切り，i 番目の bin に入るイベントの数 k_i を求める．

2. イベント数の平均および分散

$$\bar{k} = \frac{1}{N}\sum_{i=1}^{N} k_i, \quad v = \frac{1}{N}\sum_{i=1}^{N}(k_i - \bar{k})^2 \tag{4}$$

を計算する．

3. コスト関数

$$C_n(\Delta) = \frac{2\bar{k} - v}{(n\Delta)^2} \tag{5}$$

を計算する．

4. 異なる時間幅 Δ に対して 1 から 3 を繰り返し，

$$\Delta^* \equiv \arg\min_{\Delta} C_n(\Delta) \tag{6}$$

を求める．

こうして決定された最適区間幅は，データが時系列の途中で Burst 現象を示す場合には，Burst が生じている部分と生じていない部分とで異なるイベント生起率を示すようなヒストグラムを生成するため，有限の数値となる．一方，Burst 現象を示さないようなデータの場合は，時系列を通して定常状態であるため適切な区間幅を決定できず，無限大 (=時系列の長さそのもの) に近い数値となる．

2.3 critical point と burst transition

前述の最適区間幅推定の理論的背景と Hawkes Process における励起率との関係から，ある critical point

$$\alpha_c = 1 - 1/\sqrt{2} \approx 0.2929 \tag{7}$$

を境に，Δ^* の値が $+\infty$ から有限値へと移行し，スパイク時系列の挙動が Burst の生じない静的な定常状態から Burst が頻繁に生じる非定常状態へと移行することが，理論的証明およびシミュレーションによる実験によって示されている [Onaga 14]．

なお，この実証においては指数カーネルとして $f(x) = 0.1\exp(-0.1t)$ を用いている．よって，(2) 式のように係数成分をまとめて α と置くならば，$\alpha = \alpha_c \times \beta (= \alpha_c \times 0.1)$ であり，ここでいう α_c は branching ratio に対応する．

2.4 データセット

ユーザーが Web 上のコンテンツに対し自由に tag とよばれるメタ情報を投稿することでインタラクションを図る Social Tagging System を用いた Web サービスが近年増えており，内部のダイナミクスについての解析および研究も多くなされている [Taylor 16]．

今回，Social Tagging System を採用している SNS の一つである RoomClip のデータを用いて，Hawkes Process による fitting を試みた．このデータには 2012 年 4 月から 2016 年 6 月までの間に投稿された tag の ID とその投稿日時が記録されており，約 40 万種類の tag と計 880 万回の投稿が含まれている．なお，RoomClip において各 tag は同一コンテンツ (ユーザーの投稿した写真) に対して複数投稿することができ，同時に複数の tag が投稿された場合は，それらを別々に計測している．

今回はこれらのうち 4000 回以上投稿された 171 種類の tag について，それぞれの投稿日時の記録を点過程におけるイベントとして捉え，指数関数カーネルを用いた Hawkes Process による fitting を行い，最適な α, β, ρ を最尤法によってそれぞれ推定した [Ogata 81]．この際，SNS のユーザー数の増加等といった外的な要因が内部のダイナミクスに与える影響を抑えるため，2 週間ごとの総投稿数に応じて時刻を rescaling している．図 1 に RoomClip における総投稿数の推移を，図 2 に tag の投稿日時の rescaling を可視化した例を掲載する．

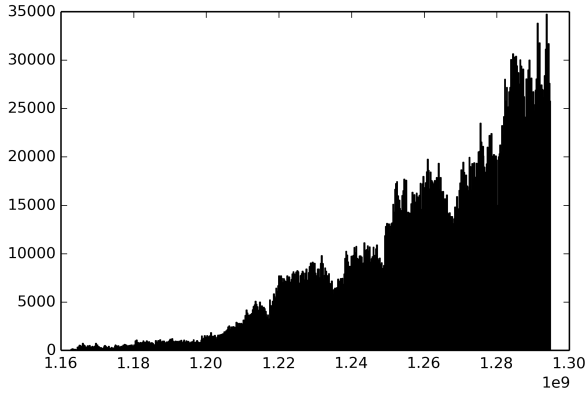

図 1: 総投稿数の推移．横軸は投稿時刻，縦軸は全体を 1000 個の bin に区切った時間幅 (約 40 時間) での tag の総投稿数を示す．

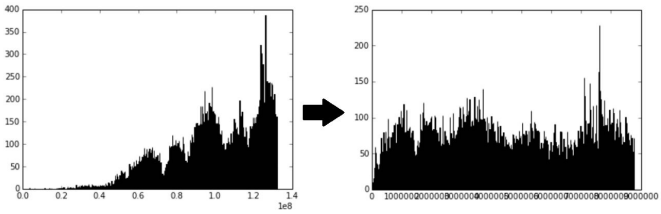

図 2: 総投稿数による各タグの投稿日時の rescaling．ある 1 つの tag の投稿時系列について，rescaling を行う前 (左) と行った後 (右) の変化を示す．図 1 の総投稿数を基にした正規化を行うと同時に，最尤推定による Hawkes Process の fitting が $\beta \approx 0.1$ で可能となるよう，全体の Time Scale の伸縮も行っている．どちらも総投稿数の場合と同様，全体を 1000 個の bin に区切った時間幅での投稿数を示している．

3. 結果

3.1 fitting の評価

Hawkes Process による fitting の評価手法としては，QQ-plot による可視化がよく用いられる．

Hawkes Process による fitting を行ったデータの時間列 $\{t_i\}$ と推定したパラメータ ρ, α, β を用いて，新たな時間列 $\{\tau_i\}$ を以下のように定義する．

$$\tau_0 = \int_0^{t_0} \rho\, ds = P(0, t_0) \tag{8}$$

$$\tau_i = \tau_{i-1} + P(t_{i-1}, t_i) \tag{9}$$

ただし，

$$P(t_{i-1}, t_i) = \int_{t_{i-1}}^{t_i} \rho\, ds + \frac{\alpha}{\beta}\left(1 - e^{-\beta(t_i - t_{i-1})}\right) A(i-1) \tag{10}$$

であり，

$$A(0) = 0 \tag{11}$$

$$A(i) = 1 + e^{-\beta(t_i - t_{i-1})} A(i-1) \tag{12}$$

このとき，Hawkes Process による fitting の当てはまりが良ければ，$\{\tau_i\}$ は Poisson Process に従い [Daley 07]，$\{\tau_i\}$ に

おける各イベントの間隔

$$\delta\tau_i = \tau_i - \tau_{i-1} \tag{13}$$

によって構成される時間列のパーセンタイル値は指数分布のそれと一致する．

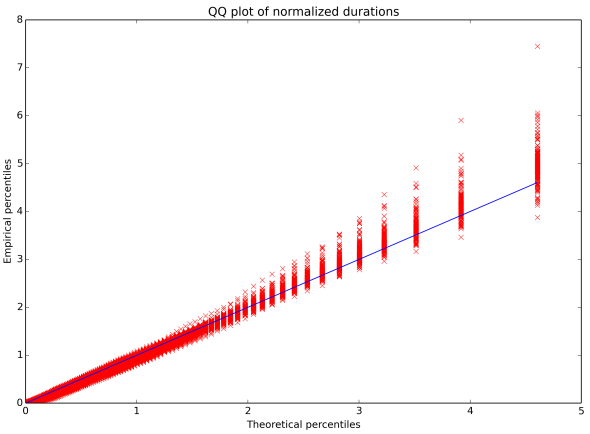

図 3: QQ-plot．横軸は理論上のパーセンタイル，縦軸は実データを用いて得られた各 tag についてのパーセンタイル値である．

これを各 tag についてプロットしたものが図 3 に示す QQ-plot である．fitting の当てはまりが良ければ，理論値と実験によって得られた結果の値が一致するので，$y = x$ の直線に沿ってプロットされる．この図から，今回の fitting が概ね良好であることが確認された．

3.2 branching ratio と最適区間幅との関係

先述の PSTH における最適区間幅の推定手法を各 tag の投稿日時の時系列データに対して適用し，求まった最適区間幅の逆数と Hawkes Process による fitting の結果得られた branching ratio(先述の α_c に対応) との関係を，図 4 に示す．

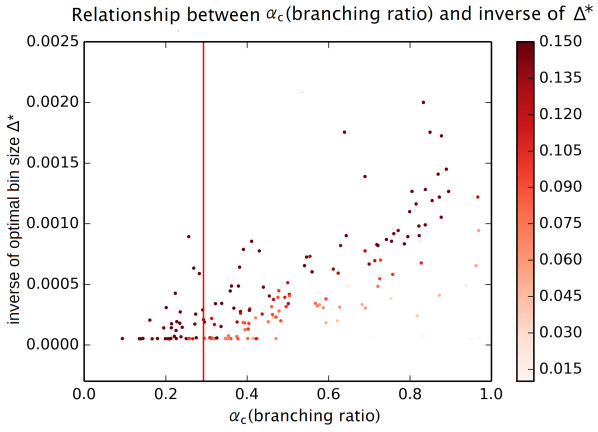

図 4: α_c(branching ratio) と最適な bin size Δ^* の逆数との関係．各点は各 tag の時系列データから得られた Hawkes Process における branching ratio (α_c に対応する) および最適区間幅の関係を示している．また，各点の色は Hawkes Process における β の値に対応する．赤線は $\alpha_c \approx 0.2929$ を示している．

最適区間幅が無限大に近い (最適区間幅の逆数が 0 に近い) 部分では，データ時系列は Burst が生じていない定常状態であり，最適区間幅が小さくなる (逆数が大きくなるにつれ) につれ動的な非定常状態となるが，実験の結果より branching ratio が大きくなるにつれ最適区間幅の逆数の値も上昇する傾向にあることが確認された．

そして，若干のばらつきはあるものの，branching ratio の値が 0.2〜0.4 の部分で定常状態と非定常状態との transition が生じており，0.4 を超えると最適区間幅は有限の値となり，必ず Burst が観測されるような状態となることが示された．これは，実データでも理論値 $\alpha_c \approx 0.2929$ の付近で transition が生じることを実証するものであると考えられる．

一方で，$\alpha_c < 0.2929$ であるにも関わらず最適区間幅が小さな値を取るようなデータも存在しており，このようなゆらぎの原因については今後細かく検証していく必要がある．

4. 議論

今回の実データを用いた Hawkes Process による fitting およびその結果と Burst 状態との関係性の検証より，理論上およびシミュレーションによって確認された critical excitability 付近において burst transition が実際に生じていることが確認された．この結果は現実のデータの統計的性質を新たに解明するものであると同時に，今回 Web データを扱ったことによる，Web における人々のインタラクションのダイナミクスの解析や Web 広告への応用にも繋がる可能性があると考えられる．

一方で，branching ratio と最適区間幅との関係においてゆらぎが大きいなどといった理論とは異なる結果も得られたことについて，2 週間ごとの総投稿数に応じた rescaling が適当であるか等を含め，実データを扱う上での処理の方法についてもう一度検討する必要があると考えられる．また，社会活動の昼夜差に起因するイベント発生率のゆらぎ等については今回の rescaling では捉えきれていない可能性が高い．このような複数の要因を考慮に入れたモデルについての検証も引き続き進めたい．

最後に，今回のモデルでは時間の経過および Web サービスのユーザーの増加に伴う総投稿数の増加を外的な要因として捉え，その影響を取り除くための rescaling を行ったが，時間の経過およびユーザーの増加に伴って Web サービスの内的構造が変化している可能性も考えられる．こうした Web の進化に関する考察についても議論していきたい．

謝辞

貴重な解析用データを提供いただいた，Tunnel 株式会社の皆様に感謝いたします．

参考文献

[Daley 07] Daley, D. J., and Vere-Jones, D.: An introduction to the theory of point processes: volume II: general theory and structure. Springer Science and Business Media (2007)

[Filimonov 12] Filimonov, V., and Sornette, D.: Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. Physical Review E, 85(5), 056108 (2012)

[Hawkes 71] Hawkes, A. G.: Point spectra of some mutually exciting point processes. Journal of the Royal

Statistical Society. Series B (Methodological), 438-443 (1971)

[Ogata 81] Ogata, Y.: ‘On Lewis’ simulation method for point processes’, IEEE Transactions on Information Theory 27(1), 2331 (1981)

[Oka 15] Oka, M., Hashimoto, Y., and Ikegami, T.: Open-ended evolution in a web system. In Late breaking papers at the European Conference on Artificial Life (p. 17) (2015)

[Onaga 14] Onaga, T., and Shinomoto, S.: Bursting transition in a linear self-exciting point process. Physical Review E, 89(4), 042817 (2014)

[Shimazaki 07] Shimazaki, H., and Shinomoto, S.: A method for selecting the bin size of a time histogram. Neural computation, 19(6), 1503-1527 (2007)

[Taylor 16] Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beslon, G., ... and McMullin, B.: Open-ended evolution: perspectives from the OEE workshop in York. Artificial Life (2016)