The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Hawkes Process を用いた Social Tagging System におけるバースト現象解析

Burst Phenomenon Analysis in Social Tagging System using Hawkes Process

江島昇太 *1 Shota Ejima　小杉太一 *1 Taichi Kosugi　岡瑞起 *1 Mizuki Oka　三宅雅矩 *2 Masanori Miyake　池上高志 *2 Takashi Ikegami

*1筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

*2東京大学大学院総合文化研究科
Graduate School of Arts and Science, The University of Tokyo

In the activities of people on WEB such as SNS, the burst phenomenon is observed. Recently, Hawkes Process is used as a method to analyze the burst phenomenon. It is known that when the branching ratio, which is an index representing the internal dynamics of Hawkes Process, exceeds a certain threshold, the event time series transits from steady state to nonstationary state where the burst phenomena is likely to occur. In this paper, we focus on social tagging system among data obtained from SNS, and analyze how branching ratio changes timewise with service growth.

1. はじめに

Social Networking Service (SNS) など, WEB 上の人々の活動においては, ある特定の話題に対する言及が急激に増えるような**バースト現象**が見られる. バースト現象は, あるイベントの発生が別のイベントの発生を引き起こすといった自己励起性という性質によって特徴づけることができる. 近年, このような WEB 上におけるバースト現象を, 自己励起型の確率的点過程の一つである Hawkes Process によってモデル化するといった試みが行われている [Oka 15].

Hawkes Process における内部のダイナミクスを表す指標である branching ratio がある一定の値を超えると, イベント時系列が静的な定常状態から, バースト現象が起こりやすい非定常状態へと相転移することが理論的及びシミュレーションによって明らかになっている [Onaga 14]. また, SNS から得られた実データに対してもこの相転移が見られることが分かっている [三宅 17][三宅 18].

SNS における特徴的なシステムとして, **Social Tagging System** が挙げられる. Social Tagging System では, ユーザがコンテンツに対して「タグ」を付与することによってコンテンツの管理や検索を容易にすることが目的とされている. 一方で, このタグはユーザが自由に付与することが出来る場合が多く, 新たなタグの発明や流行といった現象が見られる. こうした流行現象は, 強い自己励起性によってもたらされる可能性があり, ある種のバースト現象であると捉えることができる.

本研究では, SNS から得られる実データの中でも特に, Social Tagging System におけるユーザのタグ付け行為に着目した. サービスが成長するにつれて多くのタグが生み出され使用されていくが, その使用履歴にはバースト現象を呈するものが存在する. タグが付与されたコンテンツの投稿をイベント時系列とみなし, サービス初期から直近のデータを一定間隔ごとに区切った後, Hawkes Process によってモデル化することで, サービスの成長に伴って branching ratio がどのような時間的変化をしているのか, またサービスの期間内で, 非定常状態と定常状態を示す挙動がどのような比率で存在しているかを確認した.

連絡先: 江島昇太, eji@websci.cs.tsukuba.ac.jp
小杉太一, tai_kos@websci.cs.tsukuba.ac.jp

2. 解析手法

2.1 Hawkes Process

Hawkes Process は, 自己励起型の確率的点過程の一つである [Hawkes 71]. 指数関数カーネルを用いて, 強度関数 (イベントの発生率) $\lambda(t)$ が, 以下の式で表される.

$$\lambda(t) = \mu + \alpha \sum_{t_k<t} e^{-\beta(t-t_k)} \tag{1}$$

ここで, 強度とは単位時間あたりのイベントの平均発生回数を表す. α はイベント発生時の強度関数の励起率 (ジャンプ幅) を表し, β はイベント発生語の強度関数の減衰率を表す. また, μ は常に最低限保たれるベースラインとなる強度を表す.

Hawkes Process の大きな特徴は, 強度関数が過去に発生したイベントとの時間的関係によって表されており, あるイベントが起こるとさらにそれが起こりやすくなる, というような**正のフィードバック効果**が考慮されているという点である.

従来, Hawkes Process は地震の発生間隔といった自然現象の解析に対して用いられてきたほか, 神経細胞の活動や金融取引パターンの解析, 予測など, 様々な分野に応用されてきた. 近年では, SNS のおけるユーザの活動のダイナミクスを捉えるために, Hawkes Process を用いてモデル化するといった研究もなされている.

また, Hawkes Process における内部のダイナミクスを表す指標である **branching ratio** は以下のように表される.

$$n = \int_{-\infty}^{\infty} \alpha e^{-\beta t} dt = \frac{\alpha}{\beta} \tag{2}$$

branching ratio が以下に示すある一定の値を超えると, イベント時系列が静的な定常状態から, バーストが頻繁に発生する非定常状態へと相転移する比率が以下で与えられる [Onaga 14].

$$\begin{aligned} n &> 1 - 1/\sqrt{2} \\ &\approx 0.2929... \end{aligned} \tag{3}$$

すなわち, branching ratio はバースト現象を定量化する際のティッピング・ポイントとなる指標である.

2.2 最適区間幅推定

イベント時系列のバースト現象を定義する手法として, 生理学などでよく用いられる Peristimulus Time Histogram (PSTH) 作成における最適区間幅推定を用いる. 時間幅の決定に関しては, 平均二乗誤差 (MISE) 最小化の観点から, 以下の手順で求めることができる [Shimazaki 07].

1. n 回の試行による時系列の観測期間 T を幅 Δ, N 個の bin に分割し, i 番目の bin に入るイベントの個数を k_i とする.
2. bin ごとのイベント数の平均 $\overline{k}$ と分散 v

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

をそれぞれ計算する.

3. コスト関数

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

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,

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

となる Δ^* を求める.

イベント時系列が静的な定常状態である場合, この手順により求まる最適区間は無限大 (= 時系列の長さそのもの) になる. 一方, バースト現象を起こしやすい非定常状態である場合, バースト現象が発生した時点とそうでない時点とで異なるイベント発生率を示すようなヒストグラムとなるため, 有限の値となる. これを利用して, イベント時系列が定常状態か非定常状態かを判別することが出来る.

2.3 最尤法

イベント時系列に対して Hawkes Process のフィッティングを行うことで, α, β, μ を推定することができる. このパラメータ推定には, Ogata らが提唱した最尤法 [Ogata 78] がよく用いられる. 強度関数 $\lambda(t)$ から解析的に求められる対数尤度を最大化 (= 負の対数尤度を最小化) するようなパラメータ $\boldsymbol{\theta} = (\alpha, \beta, \mu)$ を求めることで, イベント時系列に対するパラメータ推定を行う.

3. 解析

3.1 実験データ

本研究では, Tunnel 株式会社によって運営されている RoomClip[*1] という部屋の写真を投稿する SNS のデータを用いる. データの取得期間は, サービス開始の 2012 年 4 月から 2016 年 6 月の約 4 年間である. 2016 年 6 月時点での RoomClip の基本データとしては, 写真数が 410,391, タグの種類数が 405,484, タグが付与された回数が 8,805,112 である.

RoomClip においても, 投稿する写真への Social Tagging System が採用されている. タグには任意の文字列を指定することができ, 一つの投稿に対して複数のタグを付与することが可能である.

*1 http://roomclip.jp/

3.2 データの選択と前処理

Hawkes Process におけるフィッティングを正しく行うためにはある程度のデータ数が必要となる. 本研究では RoomClip の全タグの中でも 4,000 回以上使用された計 171 個のタグを解析対象とした. これらのタグについて, サービスにおける何番目の投稿で対象のタグが使用されたかを, 各タグの使用を表すイベント時系列データとした. また, パラメータの時間的変化を見る方法として, 全投稿を 30,000 投稿ずつのウィンドウに分け (図 1 参照.), 各タグのウィンドウごとに Hawkes Process のフィッティングを行ない, 2.2 の 1-4 の方法を用いて, 最適な Δ^* を決定した. ただし, ウィンドウ内で 70 回以上使用されなかったデータに関してはフィッティングを行わなかった. また, 区切られたウィンドウの個数は 49 であった.

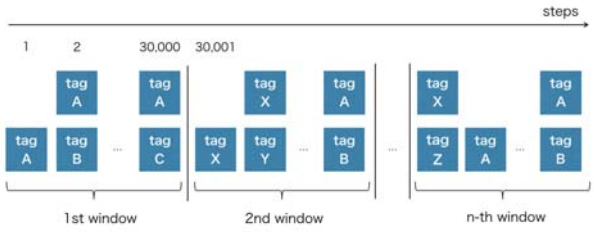


図 1: タグの投稿時系列をウィンドウに区切るイメージ図. 「そのタグが最初の投稿から何番目の投稿か」によってインデックスを振り, 30,000 投稿毎にウィンドウに区切る. 一つの投稿に対して複数のタグが付与されることもあり, その場合はそれぞれのタグに同じインデックスを振る.

3.3 結果と評価

Hawkes Process のフィッティングの評価には, QQ-plot が頻繁に用いられる. Hawkes Process によるフィッティングがうまくいっていれば QQ-plot のプロットは直線 $y = x$ 上に並ぶ. 詳細は省略するが, 本研究においては QQ-plot のプロットが大きく逸脱したものはなく, フィッティングは概ね良好であった. しかし, フィッティングがかなりうまくいっているものもあれば多少ずれているものもあり, タグごとにゆらぎが存在した.

図 2 は, 「リメイクシート」というタグに対して, 投稿数の推移 (上側) と, ウィンドウ毎に分けた投稿時系列の branching ratio の推移 (下側) を示したものである. 左側図中の赤い破線は (3) 式で示したティッピング・ポイントの値である. この「リメイクシート」というタグは, ある時点で爆発的に使用回数が増えていることが確認できる. 一方, branching ratio の値も, その爆発に合わせてティッピング・ポイントを超えた高い値を取っていることがわかり, バースト現象による相転移をうまく捉えていることがわかる.

図 3 は, 各ウィンドウごとの全タグの branching ratio の平均値をプロットした結果である. ここで, タグごとに最適な Δ^* は異なる. SNS では, サービス内のユーザ数, 投稿数が増えるに連れてユーザ間のインタラクションや他ユーザの投稿を目にする機会は増加していくため, サービスが成長するにつれてバースト現象が起こりやすくなる, すなわち branching ratio の値も上昇していくのではないかと予想したが, 平均としての結果は 0.15 程度に収まることが分かった. これに対して, 各写真への ‘ Like ’ を時系列とした branching ratio の計算では, サービスの成長とともに branching ratio が臨界的に漸近するという結果になった [Oka 17][Ikegami 17]. これは, サー

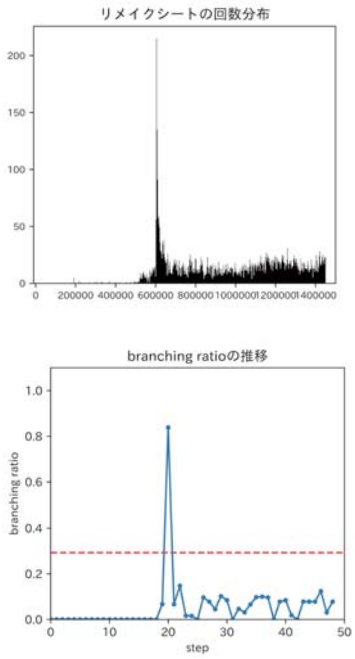


図 2: タグ「リメイクシート」の投稿数の推移 (上側) と, branching ratio のウィンドウ毎の推移. 赤い破線は (3) 式の値を示す.

ビスの成長と共にユーザ数が増え, 投稿はしないが Like は行うユーザ行動が反映されていると考えられる.

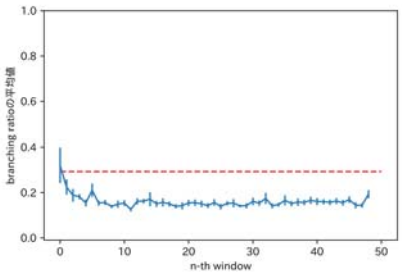


図 3: 各ウィンドウごとの全タグの branching ratio の平均値の変化. 青い線が branching ratio の平均値を示しておりエラーバーは分散を示している. 赤い破線は (3) 式のティッピング・ポイントの値である. 各ウィンドウ間でオーバーラップはさせず, 49 のウィンドウそれぞれに含まれるタグに対して計算を行っている.

図 4 は, 対象となる全タグについて, 横軸に 10 ウィンドウごとの branching ratio, 縦軸に最適区間幅 Δ^* の逆数をプロットした結果である. 時系列が定常状態であれば, 最適区間幅は無限大になるため, 逆数の値は 0 に近づく. 逆に, バースト現象が起こりやすい非定常状態では, 最適区間幅は有限の値にあんるため, 逆数の値は 0 から離れていく. 図中の赤の破線は, (3) 式で表したティッピング・ポイントの値を示しており, この破線の右側ではプロットが 0 から離れた場所にあることが分かる. これより, 実データに対しても $n > 0.2929\cdots$ となる時系列においては, 非定常状態になっており, 逆に, 破線の左側では, プロットは 0 付近に密集していることが分かり, 定常状態から非定常状態への相転移を確認することができた. 一方で, 時系列をウィンドウごとに区切ることによってサービスの開始初期から時間が経った後で branching ratio と最適区間幅の逆数との関係について大きな変化は見られなかった. つまり, サービスの成長にともなってユーザ数や投稿数が増えても, branching ratio がティッピング・ポイントを超える割合に大きな変化はなく, バースト現象の発生にはサービスの成長度合いはあまり関係ない可能性が示唆される.

表 1 に, 各タグについて branching ratio を計算することができたウィンドウに対して, (3) 式のティッピング・ポイントを超えた回数の割合を計算した結果を示す. 「クリスマス」「ハロウィン」など, 季節性のあるタグは, branching ratio の値がティッピング・ポイントの値を超える割合が非常に高いことが分かる. また, その他にも「男前」「RC 愛知」「子供と暮らす。」など, RoomClip のユーザ独自のタグも割合が比較的高い. 一方, 「 ベランダ」「収納」など, 部屋に関連する基本的なタグは割合が低いことがわかる.

表 1: branching ratio がティッピング・ポイントを超えた割合.

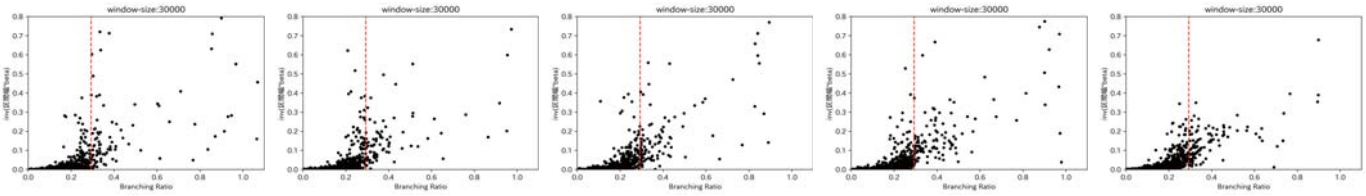
割合	タグの種類数	タグの例
10% 以下	124	ベランダ, 収納
11 - 20%	27	多肉植物, 西海岸
21 - 30%	9	男前, RC 愛知
31 - 50%	6	こどもと暮らす。
51 % 以上	5	クリスマス, ハロウィン

4. おわりに

4.1 まとめと考察

本研究では, SNS における Social Tagging System に着目し, そこから得られる実データに対して Hawkes Process によるフィッティングを行い, branching ratio の時間的変化を追うためにイベント時系列をウィンドウに区切った解析を行った. サービスの成長に伴い, バースト現象の増加, すなわち branching ratio も増加していくものと予想していたが, 結果として大きな相関は見られず, サービスの期間を通してほぼ一定であることが分かった. 一方で, 季節性のあるタグや, RoomClip 独自のタグなどは, branching ratio がティッピング・ポイントを超える割合が高いことが分かった. これは, 「部屋の写真を投稿する」という RoomClip の特性上, クリスマスなどのイベント事にはそのテーマに沿った家具やインテリアが配置された写真が多くのユーザによって局所的に投稿されるためであると推測できる.

また, 三宅らの研究 [三宅 18] においても, SNS 上の実データに対して Hawkes Process を用いて定常状態から非定常状態への相転移を捉えている. 本研究では, [三宅 18] の結果よりもばらつきが大きい結果となった. これは全てのタグに対して同じウィンドウ幅 (= 30,000 投稿) を設定したが, 実際はタグ毎に適切なウィンドウ幅が異なるためであると考えられる. ウィンドウ幅を固定すると, ウィンドウ内で有効となるサンプル数が様々であったり, また一つのバースト現象を同じウィンドウ内で囲みきれなかった場合 (バースト現象が 2 つのウィンドウにまたがってしまった場合) などが原因となり, パラメータ推定がうまく行われなかった可能性が考えられ, 今後はこのウィンドウ幅を QQ-plot による誤差が最小となるようなもの



[1] 1 - 10 ウィンドウ. [2] 11 - 20 ウィンドウ. [3] 21 - 30 ウィンドウ. [4] 31 - 40 ウィンドウ. [4] 41 - 49 ウィンドウ.

図 4: 各ウィンドウにおける branching ratio と最適区間幅の逆数の関係. それぞれのウィンドウに含まれる全タグに関してまとめて表示してある.

に設定していくなどして最適化していくことが必要であると考えている.

4.2 今後の展望

複数のイベント時系列間のダイナミクスを考慮した多次元 Hawkes Process においても, 対象としているイベント時系列のうち少なくとも一つのイベント時系列が定常状態から非定常状態へと相転移する条件の閾値は, 以下の様に表されることが知られている [Onaga 14].

$$\max_i(\boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^T\boldsymbol{\Lambda}^{-1})_{ii} > 2 \tag{7}$$

この閾値は, fluctuation condition と呼ばれる.

これに対して三宅らの計算 [三宅 18] では, ユーザ同士のフォロー・フォロワーネットワークを考え, 投稿に対する Like のやり取りに着目している. ネットワーク中の各ユーザの Like のイベント時系列に対する branchign ratio の最大成分の値は, fluctuation condition < 2 の場合は 0.29 未満であるのに対して, fluctuation condition > 2 の場合は 0.30 を超えている. 両者の fluctuation condition の差はこの値の違いによる部分が大きいと考えられる. 一方で, ネットワーク内の各ユーザに対する branching ratio の行列の各成分を見てみると, 高い非定常性示したネットワークでは非対角成分でも比較的高い値を持つものが確認できるが, 高い定常性を示したネットワークでは対角成分以外では全て低い値を取っている. この非対角成分に関する差が, Hawkes Process によって求められる fluctuation condition と最適区間幅推定によって得られる非定常性との不整合の原因となっているのではないかと考え, 特にこの不整合性の大きいいくつかのネットワークについての非対角成分の分布を比較しているが, 全体として特に差は見られていない.

本研究では今後, タグをノード, タグ同士の共起関係をエッジとするようなネットワークを考えることによって, 複数のタグ同士の関係性, 影響度を考慮したバースト現象分析を行うことができると考えている. また, ネットワーク中のエッジの有無を操作することによって, バースト現象を引き起こしたり, あるいは抑制したりすることができることがシミュレーションにより明らかになっている. 今後の実験では, タグの共起関係ネットワークにおいてこのようなエッジの操作を行うことによってバースト現象に関するシミュレーションも行いたいと考えている.

謝辞

解析に用いた貴重なデータを提供頂いた, Tunnel 株式会社の皆様に感謝致します. 本研究は JSPS 科研費 15K00420 の助成を受けたものです.

参考文献

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

[Hawkes 71] Hawkes, A. G.: Point spectra of some mutually exciting point processes. Journal of the Royal Statistical Society. Series B (Method- ological), pp. 438-443. (1971)

[Onaga 14] Onaga, T. and Shinomoto, S.: Bursting transition in a linear self-exciting point process. Physical Review E, 89(3), 042817. (2014)

[Shimazaki 07] Shimazaki, H., and Shinomoto, S.: In A Method for Selecting the Bin Size of a Time Histogram, Neural computation, 19(6), pp. 1503-1527. (2007)

[Ogata 78] Ogata, Y.: The asymptotic behaviour of maximum likelihood estimators for station- ary point processes. Annals of the Institute of Statistical Mathematics, 30(1), pp. 243-261. (1978)

[Oka 17] Oka Mizuki, Hashimoto Yasuhiro, Ikegami Takashi: Understanding evolutionary dynamics in online social networks. In Proc. of IEEE ALIFE 2017, DOI 10.1109/SSCI.2017.8280796, pp. 1-5. (2017)

[Ikegami 17] Takashi Ikegami, Mizuki Oka, Yasuhiro Hashimoto, Yoichi Mototake: Life as an Emergent Phenomenon, Studies From Large-Scale Boid Simulation and Web Data, Philosophical Transactions A of the Royal Society, 375.20160351. (2017)

[三宅 17] 三雅雅矩, 池上高志, 岡瑞起, 橋本康弘: Hawkes Process と最適区間幅推定を用いた Web Data の解析. 人工知能学会全国大会 (2017)

[三宅 18] 三宅雅矩: Hawkes 過程を用いた Web データにおける相転移現象の解析. 東京大学大学院総合文化研究科広域科学専攻, 平成 29 年度修士論文 (2018)