3K4-3in2

大規模ソーシャルデータを用いた話題継続性のモデリング

Index of topics' survival calculated from large-scale social data

山田健太 [1,2*] 玉岡諒 [1] 和泉潔 [1]
Kenta Yamada[1] Ryo Tamaoka[1] Kiyoshi Izumi[1]

[1] 東京大学工学系研究科
[1] Graduate School of Engineering, The University of Tokyo
[2] 科学技術振興機構 PRESTO
[2] PRESTO, JST

Abstract: We show results of analyzing big blog data and introduce an index which is valuable to forecast future popularity, especially comedian popularity after the winning or coming second in a comedian contest. There are some popular comedian contests in Japan such as the M- 1 grand prix. We can universally observe clear peak and power law decaying in the number of blog entries including a champion and vicechampion name (34 samples) after contests as well as the cases in which the number of entries including the event's name follows power function after the events such as Christmas in the previous study[1]. We fitted the number of entries including the comedian's name using five days data after the contest by a power function and cumulated differences between fitted line and actual data from 6 to 12 days after the contest. We found that this index of cumulative differences has a good predictive capability for the number of future (11months later) entries about the comedian.

1 はじめに

インターネットの普及により，様々な人々がブログや Twitter に様々な話題を書き込むようになった．図. 1は 2015 年 9 月 1 日から 2015 年 12 月 31 日まで「ラグビー」「五郎丸」「ルーティーン」が 1 日あたりブログに何件書き込まれたかを表している．9 月 20 日 (日本時間) の南アフリカ戦に勝利したことにより，今までは 1 日数百件程度だった「ラグビー」の書き込み数が一気に７０００件程度書き込まれた．その後も試合があるごとに話題になっている様子がわかる．ラグビーの盛り上がり並行して，五郎丸選手の書き込みや，五郎丸選手がプレースキックの際に行う一連の動作「ルーティーン」の書き込みが増え，ラグビーやその周辺の話題が盛り上がっていった様子を定量的に捉えることができる．

このように，ブログや Twitter に代表されるソーシャルメディアに関する大規模データの解析が可能となり，これらのデータを解析することにより，現在の世間の話題を把握すること (ナウキャスト) や将来ある話題がどれくらい盛り上がる，もしくは衰退するのかを予測すること (フォアキャスト) は，商品の生産などを戦略

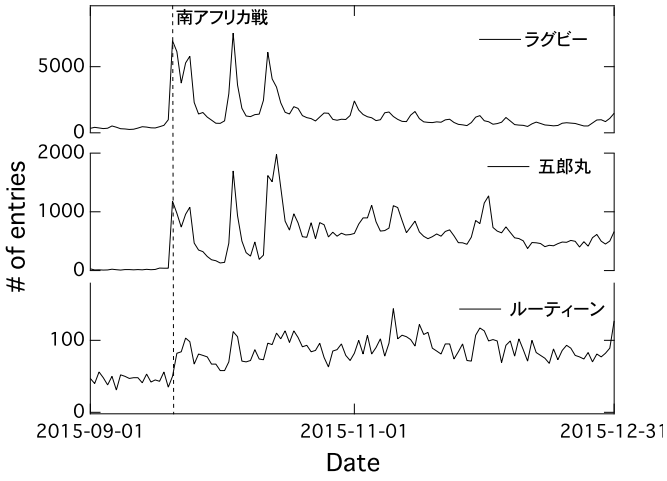

図 1: 「ラグビー」,「五郎丸」,「ルーティーン」の 1 日あたりの書き込み数．2015 年 9 月 1 日から 2015 年 12 月 31 日まで．点線は南アフリカ戦のあった 9 月 20 日 (日本時間)．

*連絡先：東京大学
東京都文京区本郷 7-3-1 工学系研究科システム創成学専攻
E-mail:yamada@sys.t.u-tokyo.ac.jp

的に行うためにも重要であり企業からのニーズも非常に高い．

しかし，任意の単語に対して未来の書き込み数の予測を行うことは容易ではない．つまり，世間における話題の変化は，人々のコミュニケーションという非線形な相互作用が背後にあるため非常に複雑であり，すべての単語のダイナミクスを表すモデルを構築することは困難である．一方，限定した範囲では，ある程度予測可能な事象もある．例えば，訃報後の有名人の名前の書き込み数は，訃報を伝えるニュースの効果によって一気に書き込み数が増え，その後，べき関数的に減衰することが知られている [1][2]．つまり，突発的なニュースの後の減衰はべき関数のモデルを用いればある程度予測が可能である．

同様に，M-1 グランプリなどで優勝したお笑い芸人の名前も優勝後に一気に跳ね上がり，その後べき関数で減衰し，訃報後の有名人の書き込み数と共通した性質を持つ．本発表では，このべき関数で減衰する特徴を用いて，M-1 グランプリなどのコンテスト 2 週間後のデータから，優勝したお笑い芸人の１１ヶ月後の書き込み頻度と相関の強い特徴量を提案し予測可能性があることを示す．

2　話題継続指数の提案

本研究では，ホットリンク社提供の口コミ係長 [1] から，ブログにあるお笑い芸人の名前が 1 日あたり何件書き込まれているか ($w_j(t), j$: お笑い芸人名のインデックス) を取得した．口コミ係長は，単語と期間を与えると約 3000 万ブロガーが投稿した，約 25 億記事のブログデータベースを検索し 2006 年 11 月から任意の単語書き込み数を日次で調べることが可能である．また，1 日あたりのブログ総数 ($W(t)$) を用いて規格化を行う．

$$R_j(t) = \frac{w_j(t)}{W(t)} \tag{1}$$

$R_j(t)$ は，ブログ中にお笑い芸人名 j が何割含まれるかを表し，これを書き込み頻度として用いる．この規格化によって，ブログ自体の流行り廃りや週の周期性 (ブログは土日に書き込まれやすい) などの影響を排除することができる．

図.2 は 2009 年の M-1 グランプリで準優勝したお笑い芸人「笑い飯」の書き込み頻度 $R_j(t)$ の時系列である．$\tau = 0$ が M-1 グランプリ当日であり，急激に書き込み数が上昇し，その後急速に減衰している．実線は，コンテスト後 5 日間のデータを用いてべき関数 ($\hat{R}_j(\tau) = \alpha\tau^{-\beta}$) でフィッティングした結果である．コンテスト直後のインパクトの大きさを表す係数 α とべき減衰の傾きを特徴付ける β は，それぞれ $\alpha = 0.10, \beta = 1.31$ である．たった 5 日間のデータのみを用いてフィッティングを行っているが，その後もフィッティングしたべき関数で減衰していることが確認できる．

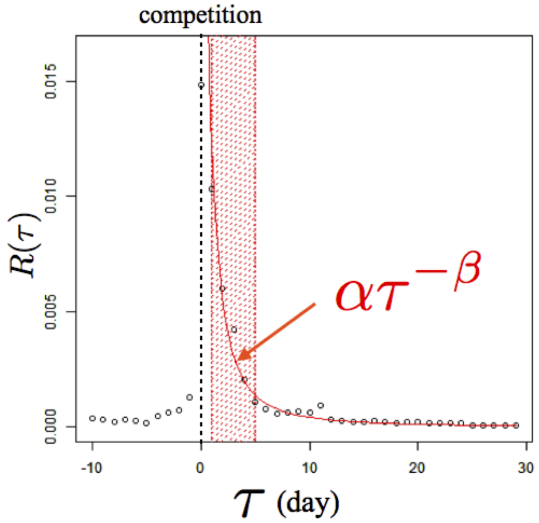

図 2: 2009 年の M-1 グランプリで準優勝したお笑い芸人「笑い飯」の書き込み頻度 (○)．実線は，コンテスト後 5 日間のデータを用いてべき関数でフィッティングした結果．

一方，図.3 は 2008 年の M-1 グランプリで優勝した「NON STYLE」の場合である．「笑い飯」同様，M-1 グランプリ当日に急激に上昇しその後減衰しているが，M-1 グランプリ後 5 日間のデータを用いてフィッティングしたべき関数には，6 日目以降従わず減衰が止まっているように見える．つまり，「NON STYLE」の場合は単純なべき関数のモデルでは将来の書き込み数を推定することはできない．

この両者の差異に着目し，以下の話題継続性の指数 G_j を提案し，この G_j が大きいほど減衰が下げ止まる傾向があり，数ヶ月の書き込み頻度が高い傾向があることを示す．

$$G_j = \sum_{\tau=6}^{12} \left(R_j(\tau) - \hat{R}_j(\tau) \right) \tag{2}$$

また，話題継続性の指数 G_j のライバル指数として，M-1 グランプリ前 12 日間の書き込み頻度の和 (B_j) と M-1 グランプリ後 12 日間の書き込み頻度の和 (A_j) の 2 つを挙げる．

$$B_j = \sum_{\tau=-12}^{-1} R_j(\tau) \tag{3}$$

[1] https://www.hottolink.co.jp/service/kakaricho

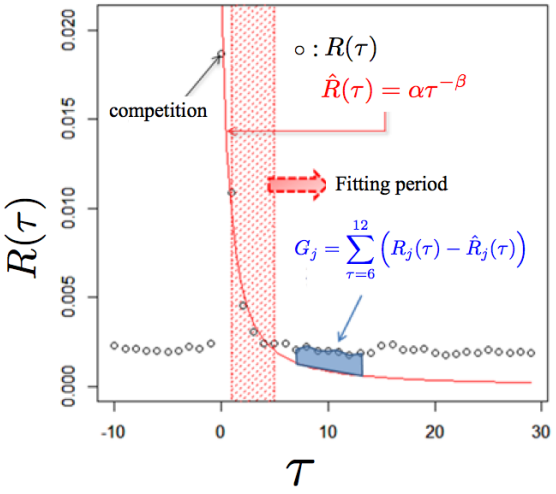

図 3: 2008 年の M-1 グランプリで準優勝したお笑い芸人「NON STYLE」の書き込み頻度 (○). 実線は，コンテスト後 5 日間のデータを用いてべき関数でフィッティングした結果.

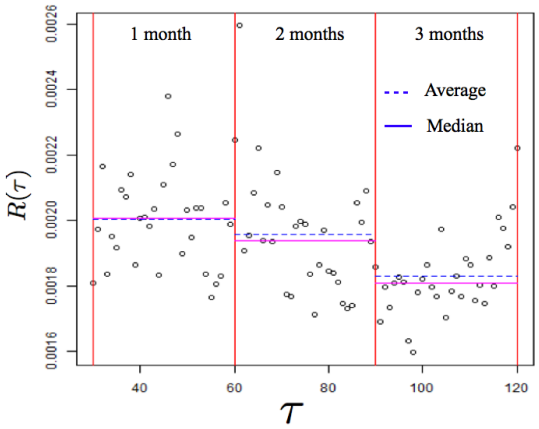

図 4: コンテスト 30〜120 日後の書き込み頻度．点線はボックス内の平均値，実線は中央値をそれぞれ表し，徐々に減衰している様子が確認できる.

$$A_j = \sum_{\tau=1}^{12} R_j(\tau) \tag{4}$$

3 話題継続指数と書き込み頻度の相関

2 節で提案した 3 つの指数 (G_j, B_j, A_j) と図.4 に示したコンテスト後の書き込み頻度の中央値との相関を，M-1 グランプリ (2006〜20010 年)，R-1 グランプリ (2006〜20012 年)，キングオブコント (2008〜20012 年) で優勝，準優勝した 3 4 組を対象に行った.

図.5 は話題継続指数 G_j(横軸) と 1 ヶ月後の書き込み頻度の中央値 (縦軸) の散布図であり，一つの点が 1 組の芸人を表す．話題継続指数 G_j が高いほど，1 ヶ月後の書き込み頻度の中央値も高いという，強い正の相関が見られ，決定係数は 0.90 である．つまり，コンテスト後 2 週間のデータを用いて計算した話題継続指数 G_j から 1 ヶ月後 (コンテストから 30-60 日) の書き込み頻度をおおよそ推定できることを意味する.

図.6 の青線は，話題継続指数 G_j と kヶ月後の書き込み頻度の中央値の間の決定係数を 1ヶ月後から 11ヶ月後まで計算した結果である．1 ヶ月後は図.5 に示した通り 0.90 であり，11ヶ月後も 0.80 程度の高い相関を示している．一方，ライバル指数として挙げた B_j と A_j は G_j と比較すると明らか低く，コンテスト後の書き込み頻度との相関が弱いことが分かる.

なぜ，G_j が大きいほど下げ止まる傾向があるかの理

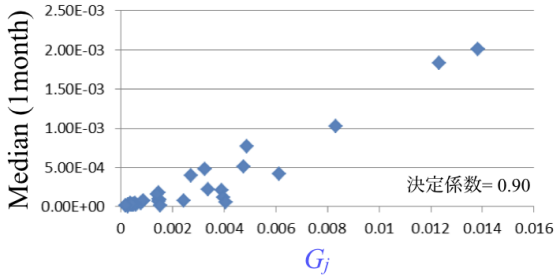

図 5: 話題継続指数 G_j(横軸) と 1 ヶ月後の書き込み頻度の中央値 (縦軸) の散布図．決定係数は 0.90 である.

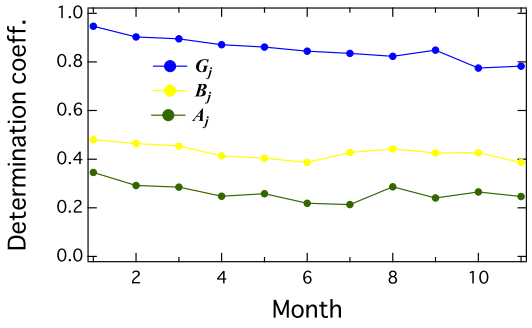

図 6: 話題継続指数 (G_j)，ライバル指数 (B_j, A_j) と kヶ月後の書き込み頻度の中央値の決定係数の推移.

由は現在解析中であるが，例えば以下のような理由が考えられる．コンテスト後５日間は，M-1 に関する話題が大半を占めると考えられる．しかし，１週間程度たった後，優勝，準優勝した芸人に対して話題が M-1 しかない場合は，図.2 の「笑い飯」の場合のようにべき関数の減衰曲線に乗り続け下げ止まらない．一方，M-1 以外の話題もある場合には，図.3 の「NON STYLE」場合のように下げ止まる傾向があると考えられる．

4　まとめ

大規模ブログデータを用いて，M-1 グランプリなどのコンテスト後に，優勝，準優勝した芸人の名前が急激に上昇しその後べき関数で減衰することを確認した．この減衰が下げ止まるかどうかは，話題継続指数 G_j によって特徴づけられることを示した．この特徴量は非常に簡単な計算から算出されるが，コンテスト後約２週間のデータから 11ヶ月後の書き込み数を推定することが可能である．

謝辞

本解析を行うにあたり，株式会社ホットリンクの口コミ係長のデータを利用させていただきました．貴重なデータを使用させてくださり心より感謝いたします．

参考文献

[1] Sano Y, Yamada K, Watanabe H, Takayasu H and Takayasu M: Empirical analysis of collective human behavior for extraordinary events in the blogosphere, *Phys. Rev. E*, 87 012805 (2013)

[2] 山田　健太, 佐野　幸恵, 高安　秀樹, 高安　美佐子: ソーシャルネットワークの力学－ブログ解析からシステム/制御/情報, Vol.56, No.10, pp536-541 (2012)