3I4-OS-02b-2

ソーシャルメディアの書き込み数における揺らぎ

Number Fluctuation of Word Appearances in Social Media

佐野 幸恵 *1 Yukie Sano
高安 秀樹 *2*3*4 Hideki Takayasu
高安 美佐子 *4 Misako Takayasu

*1筑波大学 University of Tsukuba
*2ソニーコンピュータサイエンス研究所 Sony Computer Science Laboratories, Inc.
*3明治大学 Meiji University
*4東京工業大学 Tokyo Institute of Technology

To examine number fluctuations in online social media, we focus on daily appearances of adjectives, adnominal particles and conjunctions in Japanese blog space. We show that there is a non-trivial fluctuation scaling law between standard deviation and mean. When the number of appearances gets larger, the standard deviation grows linearly. Random Posting (RP) model, that is inspired by Random Diffusion model, is introduced to reproduce the observed scaling law. For possible applications, we show one example of topic detection and we confirm that detected peaks have external events.

1. はじめに

ソーシャルメディアには様々な単語が溢れており，出現頻度や検索回数が急に増加した単語から，有益な情報を抽出することは研究が進んでいる．さらに「バースト (burst)」といった現象も定義され，そのモデル化には，単純なポアソン過程が用いられてきた [Kleinberg 03, 藤木 04].

しかし，平均的な出現頻度が変わらない単語の場合，その出現頻度はどの程度の揺らぎをもっているのかに関しては，あまり注目されてこなかった．われわれは，これまでに日本語ブログにおいて副詞や接続詞といった単語の出現頻度が，単純なポアソン過程だけでは説明できず，揺らぎのスケーリング則が存在することを明らかにしてきた [Sano 09].

そこで，本発表では揺らぎのスケーリング則について，対象単語数を増やしてデータを収集した結果を紹介する．さらに，揺らぎのスケーリング則を再現するランダム投稿モデルについて説明し，最後に応用例を示す.

2. データと前処理

ここでは，揺らぎのスケーリング則を導入するため，日本語のブログ空間で「日常的に使われる単語（以下，日常語と呼ぶ）の出現頻度時系列」を定義し，データを網羅的に収集し，行った前処理について説明する．

2.1 データ

本研究では，株式会社ホットリンクの所有するデータベースに保存されたブログデータを用いた．ホットリンクでは，20 を超える日本の主要ブログプロバイダーへ書き込まれた公開データを収集し，データベース化して，検索ツールとして有料で顧客に提供している．本研究では，データベースにアクセスして得られる，検索単語を含むブログ記事数（以下，書き込み数と呼ぶ）の 1 日刻みの出現頻度時系列を扱う．

検索単語を含む書き込み数は，データベースに保存されている記事の本文中か記事の表題に検索単語を一度でも含むかどうかで決定している．そのため，同一記事が複数回，検索単語を含んでいても，1 と数える．他方，同一ブロガーが，同日に，検索単語 を含む記事を複数投稿した場合は，別々に検索対象となるため，複数に数え上げる．

はじめに，日常語の候補となる単語の出現頻度時系列を網羅的に収集した．日常語の候補となる単語 は，ニュースや，季節性変動の外的要因に影響を受けにくいものが望ましい．そこで，形態素解析器 MeCab に標準仕様で付属している IPA 辞書 *1 の単語の中から，形容詞，連体詞，接続詞を抽出した．

「美しい」「青い」などの形容詞の場合は，原型部分だけを対象とした 1768 単語,「由々しき」「あくる」などの連体詞は 135 単語,「すると」「しかし」などの接続詞は 171 単語を検索対象とした．データの取集期間は，2006 年 11 月 1 日から，2010 年 6 月 9 日までの 1317 日間とした.

2.2 前処理

はじめに，データはホットリンク社のスパムフィルタ（「中」レベル）を用いることで，明らかなスパムブログを排除している．結果としてはフィルタを使わない場合と比較して，約 15%の記事が取り除かれている．

次に，ブログというソーシャルメディア使用者の増減の影響を差し引くため，投稿された，単語によらない全記事数 $w(t)$ による規格化を行った．ここでは単純に投稿される単語 k を含む記事数の時系列を $w^{(k)}(t)$ とし，$w^{(k)}(t)/w(t)$ とした.

最後に，定常性を確認するため，単位根検定の一つである拡張 Dickey-Fuller(ADF) 検定 [Said 84] を用いて，定常と判定されたもののみに絞り込み，最終的に形容詞 1749 単語，連体詞 132 単語，接続詞 154 単語を日常語として，本研究の解析対象とした.

3. 揺らぎのスケーリング則

ここでは，前章で集めて前処理を行ったデータに対し，揺らぎを時系列の標準偏差として定義し，計算して得られた結果と，それを再現するモデルについて説明する．

連絡先: 佐野 幸恵，筑波大学システム情報系社会工学域，〒 305-0817 茨城県つくば市天王台 1-1-1, sano@sk.tuskuba.ac.jp

*1 http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz (Accessed:2012.10.31)

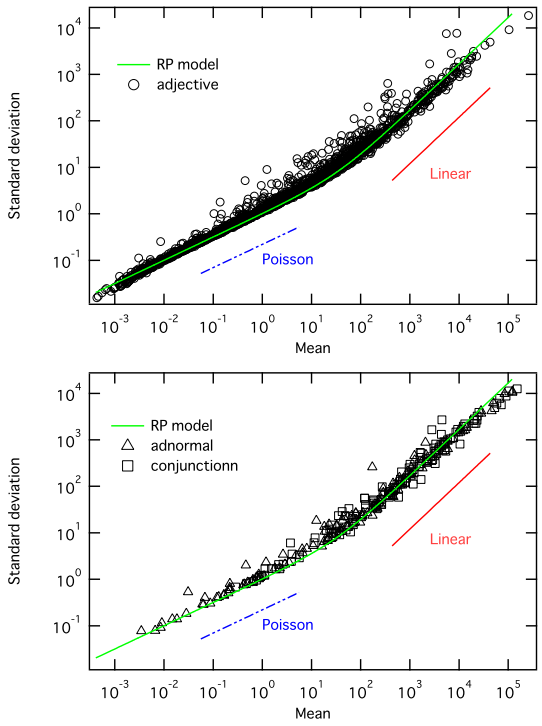

図 1: 平均値に対する標準偏差のスケーリング則 [Sano 09, Sano 13]. 形容詞 1749 単語（◯），連体詞 135 単語 (△) と接続詞 171 単語 (□). 緑の実線が RP モデルの理論解に対応する.

3.1 日本語のブログ空間での結果

単語 k を変えながら，出現頻度の平均値 $\langle w^{(k)} \rangle$ と標準偏差 $\sigma^{(k)}$ を散布図にしたのが図 1 である．出現頻度の低いエリアでは，ポアソン過程を仮定した場合の $\sigma^{(k)} = \sqrt{\langle w^{(k)} \rangle}$ を確認できる (図 1 の破線)．しかし，$\langle w^{(k)} \rangle$ が大きくなると，$\sigma^{(k)}$ が $\langle w^{(k)} \rangle$ に対して線形に増加していることがわかる (図 1 の赤の実線)．ある変数の大きさを変化させたときの，興味ある量の変換関係をスケーリングと言うが，図 1 より，平均値に対する標準偏差の非自明なスケーリング則を確認できる．

このような平均値に対する標準偏差のスケーリング則はテイラーのスケーリング則 (Taylor's scaling law)[Taylor 61] と呼ばれ，ブログ空間だけではなく，土壌中に存在する昆虫の数の分布や，河川の流量の時系列にも観測されている [Menezes 04]. 特に，前者の時間変数を伴わない，ある集合についての分布の揺らぎの関係は Ensemble Fluctuation Scaling(EFS)，後者の時系列に関する揺らぎの関係は Temporal Fluctuation Scaling(TFS) と呼ばれている [Eisler 08]. 今回観測された，ブログ空間における単語出現頻度における揺らぎはこの TFS の一種であると言える.

3.2 ランダム投稿モデル

TFS を再現するモデルとして導入されたものにランダム拡散 (Random Diffusion; RD) モデルがある [Meloni 08]. RD モデルでは，任意のネットワーク上を互いに無相関に移動するランダムウォーカーのモデルとして導入された．RD モデルは各ノード上を通過するランダムウォーカー数が作り出す時系列の揺らぎに，TFS を再現できる．しかし，その本質はネットワークではなく，観測しているネットワーク全体（系全体）に存在している，ランダムウォーカー数が揺らぐことである.

RD モデルでは，インターネット上のパケットの流量や河川の流量に関してはよい解釈を与えられる．しかし，われわれが観測した，ブログ空間における単語の出現頻度に関しては，よい解釈を与えられない．例えば，ブログ空間における単語の出現頻度と，RD モデルにおけるネットワークやランダムウォーカーとを対応づけることができない.

そこで，われわれは，ブログ空間での単語出現頻度に特化したランダム投稿モデル (Random Posting; RP) モデルを考えた [Sano 09]．RP モデルにおいては，ブログ投稿者の行動に 2 つの確率過程を導入する.

- 記事を投稿するか否か
- 投稿した場合，投稿記事に単語 k を含むか否か

RP モデルでも，各ブロガー間は無相関であると仮定し，さらに同一ブロガーの中でも，上記 2 つの行動間は無相関であると仮定する．さらに，単語 k は他の単語との共起関係は考えず，過去の自分の書き込み回数からも無相関だとすると，単語 k を含むブログの期待値 $\langle w^{(k)} \rangle$ は，単語 k の投稿確率 $p^{(k)}$ と，定数である全投稿数 w に比例し，以下で与えられる.

$$\langle w^{(k)} \rangle = p^{(k)} w \tag{1}$$

その結果，ブロガーの行動に，記事を投稿するか否かの揺らぎを考量しない場合，単語 k が n 回書き込まれる確率は，単純に期待値 $\langle w^{(k)} \rangle$ をパラメータとしたポアソン過程で与えられる.

$$P^{(k)}(n; \langle w^{(k)} \rangle) = \mathrm{e}^{-\langle w^{(k)} \rangle} \frac{\langle w^{(k)} \rangle^n}{n!} = \mathrm{e}^{-\left(p^{(k)} w\right)} \frac{\left(p^{(k)} w\right)^n}{n!} \tag{2}$$

次に，ブロガーの行動に，記事を投稿するか否かの揺らぎを考慮する場合を考える．これは，単語によらない記事数 $w(t)$ が時刻 t によって変化することと同等であり，以下の式で与えられる.

$$P^{(k)}(n) = \frac{1}{2\delta + 1} \sum_{m=-\delta}^{\delta} \mathrm{e}^{-\left[p^{(k)}(\langle w \rangle + m)\right]} \frac{\left[p^{(k)}(\langle w \rangle + m)\right]^n}{n!} \tag{3}$$

ここでは簡単のため，$w(t)$ は $[\langle w \rangle - \delta, \langle w \rangle + \delta]$ の範囲で一様分布で揺らぐ場合を考えている．単語 k が n 回書き込まれる確率は，$w(t)$ が $[\langle w \rangle - \delta, \langle w \rangle + \delta]$ の場合を等確率で足し上げれば良い.

(3) 式より，二次のモーメントは

$$\langle w^{(k)2} \rangle = \sum_{n=0}^{\infty} n^2 P^{(k)}(n) = \langle w^{(k)} \rangle^2 \left[1 + \frac{\delta(\delta+1)}{3\langle w \rangle^2}\right] + \langle w^{(k)} \rangle \tag{4}$$

となり，$\delta \simeq \delta + 1$ とすると

$$\sigma^{(k)2} = \langle w^{(k)2} \rangle - \langle w^{(k)} \rangle^2 = \langle w^{(k)} \rangle \left[1 + \frac{\langle w^{(k)} \rangle}{3} \left(\frac{\delta}{\langle w \rangle}\right)^2\right] \tag{5}$$

となる．式 (5) 式の $\delta = 0$ のとき，$\sigma^{(k)2} = \langle w^{(k)} \rangle$ となり，ポアソン過程において平均値が分散に等しいことと一致する．平均的な全投稿数 $\langle w \rangle$ に対し，揺らぎ δ が無視できなくなるとき，式 (5) において右辺第二項が支配的になり，標準偏差は平均値に対して線形に増加することが示される.

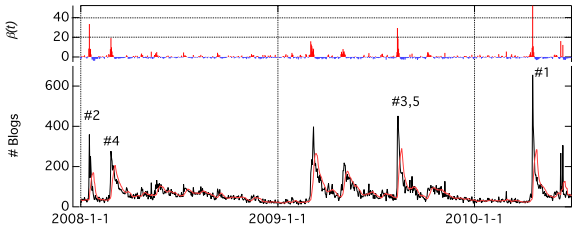

図 2: 「生茶」という商品名の書き込み時系列．下段の黒実線が全投稿数で規格化したブログ数 $w^{(\text{生茶})}(t)$，赤実線が過去 7 日間の移動平均値 $\overline{w^{(\text{生茶})}}(t)$．上段が実際のブログ数から見積もった揺らぎから逸脱した割合 $\beta(t)$ で $\beta(t) > 0$ であれば赤，$\beta(t) \leq 0$ であれば青で示している．

すなわち，全数の揺らぎ δ の大きさと全投稿数 $\langle w \rangle$ の比 $\delta/\langle w \rangle$ に対して $\langle w^{(k)} \rangle$ が大きいかどうかで，式 (5) の右辺第一項が支配的か，第二項が支配的かが決まる．(5) 式を実際のデータに重ねて示したものが，図 1 の緑の実線である．ここでは $\delta/\langle w \rangle = 0.29$ となっている．

ここまでの計算は，$w(t)$ の揺らぎの分布に $[\langle w \rangle - \delta, \langle w \rangle + \delta]$ の一様分布を仮定したが，$p^{(k)}$ が非常に小さい場合と，大きい場合で分けることで，任意の揺らぎの場合に拡張しても同等の結果が得られる [Sano 13]．この場合，平均値と標準偏差のスケーリング則は，

$$\sigma^{(k)2} = \langle w^{(k)} \rangle \left[1 + \left(\frac{\delta}{\langle w \rangle} \right)^2 \langle w^{(k)} \rangle \right] \tag{6}$$

となる．

4. 応用例 - 異常値の検出 -

式 (5) と式 (6) の結果より，平均的な全投稿数の値 $\langle w \rangle$ に対する揺らぎ δ の比 $\delta/\langle w \rangle$ だけで，平均値 $\langle w^{(k)} \rangle$ から標準偏差 $\sigma^{(k)}$ を導けることが明らかになった．そこで，様々な時系列に対して，式 (6) を適用し，許容される揺らぎから大きく逸脱した値を「異常値」としてとらえることが可能となる．

4.1 検出方法

異常値の検出には，予測される書き込み数とその揺らぎを用いる．ここでは予測される書き込み数は過去 n 日間の移動平均値 $\overline{w^{(k)}}(t)$ を用いる．

$$\overline{w^{(k)}}(t) = \frac{1}{n} \sum_{m=1}^{n} w^{(k)}(t-m) \tag{7}$$

$\overline{w^{(k)}}(t)$ に対して，式 (6) より，標準偏差は以下で予想することができる．

$$\sigma^{(k)}(t) = \sqrt{\overline{w^{(k)}}(t) \left[1 + \left(\frac{\delta}{\langle w \rangle} \right)^2 \overline{w^{(k)}}(t) \right]} \tag{8}$$

最後に，実際の書き込み数 $w^{(k)}(t)$ の揺らぎからの逸脱度 $\beta(t)$ は以下によって定義する．

$$\beta(t) = \frac{w^{(k)}(t) - \overline{w^{(k)}(t)}}{\sigma^{(k)}(t)} \tag{9}$$

表 1: 図 2 の「生茶」の書き込み時系列におけるピークの抽出例．$\beta(t)$ の大きかった上位 5 件．

#	日付	イベント	$\beta(t)$	$w^{(\text{生茶})}(t)$
1	2010.04.19	新 CM1	52.2	763
2	2008.01.17	新 CM2	33.5	261
3	2009.08.11	キャンペーン 1	29.5	299
4	2008.02.26	キャンペーン 2	19.2	216
5	2008.08.12	キャンペーン 1	18.8	339

4.2 検出結果

図 2 は「生茶」という商品名に注目し，下段に実際のブログ数 $w^{(\text{生茶})}(t)$ と，$n = 7$ 日間の移動平均値 $\overline{w^{(\text{生茶})}}(t)$，上段に逸脱度 $\beta(t)$ を表示した．$n = 7$ とした理由は，単語ごとに持ちうる週の周期性の影響を取り除くためである．表 1 には，$\beta(t)$ の大きい順に日付を抽出し，商品の公式ウェブサイトなどで確認して対応付けした．

逸脱度 $\beta(t)$ は 2010 年 4 月 19 日に最大値をとり，この日付には人気のあるタレントによる新 CM 放送開始という話題があり，明らかな外的要因を対応づけることができる．その他の日付にも，それぞれ外的要因を確認できる (表 1)．

5. まとめ

本研究では，ソーシャルメディアの中でも日本語のブログ空間に，日常的に出現する単語の日次出現頻度に着目し，その時系列の揺らぎについて報告した．特に時系列が定常と判定された形容詞，連体詞，接続詞について調べたところ，平均値と標準偏差の間に非自明なスケーリング則が存在していることを指摘した．このようなスケーリング則は，日本語のブログ空間における単語の出現頻度だけではなく，株価の変動時系列や，河川の流量時系列などにも報告され，テイラーのスケーリング則と呼ばれている．

そこで，テイラーのスケーリング則を再現する，ネットワーク上のランダムウォーカーを使ったランダム拡散 (RD) モデルを基にして，ネットワーク構造を考慮しない，ブログの投稿行動に特化したランダム投稿 (RP) モデルを紹介した．RP モデルは単純な 2 つの確率過程の足しあわせであり，その解析解として平均値から標準偏差を導出することができる．

最後に，実際の商品名を用いて，RP モデルを応用することで容易に時系列中の異常値を定量化して抽出する例を示した．抽出した異常値を確認したところ，それぞれに商品の CM やキャンペーンの影響を確認することができた．

本研究でモデル化に用いたデータは，形容詞，連体詞，接続詞というトレンドなどの影響を受けにくい単語である．しかしながら，名詞やトレンドのある単語の場合は投稿者間や単語間の相関の影響は無視できないことが考えられる．単語の種類を広げ，どの程度まで RP モデルが異常値検出に対応可能であるかについては，今後詳細に調査を行うことが必要である．

参考文献

[Eisler 08] Eisler, Z., Bartos, I., and Kertész, J.: Fluctuation Scaling in Complex Systems: Taylor's Law and Beyond, *Advances in Physics*, Vol. 57, No. 1, pp. 89–142 (2008)

[Kleinberg 03] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Data Mining and Knowledge Discovery*, Vol. 7, No. 4, pp. 373–397 (2003)

[Meloni 08] Meloni, S., Gómez-Gardeñes, J., Latora, V., and Moreno, Y.: Scaling Breakdown in Flow Fluctuations on Complex Networks, *Physical Review Letters*, Vol. 100, No. 20, p. 208701 (2008)

[Menezes 04] Menezes, Argollo de M. and Barabási, A.-L.: Fluctuations in Network Dynamics, *Physical Review Letters*, Vol. 92, No. 2, p. 028701 (2004)

[Said 84] Said, S. E. and Dickey, D. A.: Testing for Unit Roots in Autoregressive-moving Average Models of Unknown Order, *Biometrika*, Vol. 71, No. 3, p. 599 (1984)

[Sano 09] Sano, Y., Kaski, K. K., and Takayasu, M.: Statistics of Collective Human Behaviors Observed in Blog Entries, in *Proceedings of the 9th Asia-Pacific Complex Systems Conference*, pp. 195–198 (2009)

[Sano 13] Sano, Y., Yamada, K., Watanabe, H., Takayasu, H., and Takayasu, M.: Empirical Analysis of Collective Human Behavior for Extraordinary Events in the Blogosphere, *Physical Review E*, Vol. 87, No. 1, p. 012805 (2013)

[Taylor 61] Taylor, L.: Aggregation, Variance and the Mean, *Nature*, Vol. 189, No. 4766, pp. 732–735 (1961)

[藤木 04] 藤木 稔明, 南野 朋之, 鈴木 泰裕, 奥村 学.: document stream における burst の発見, 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 23, pp. 85–92 (2004)