

日本語ブログにおける単語の書き込み頻度時系列の統計的性質 とその社会的イベント計量への応用

情報・システム研究機構 社会データ構造化センター, 統計数理研究所 リスク解析戦略研究センター
渡邊隼史

基礎数理[1]: 筑波大学システム情報系 佐野幸恵, ソニーCSL 高安秀樹, 東京工業大学情報理工学院 高安美佐子

応用[2]: 株式会社ホットリンク 小森あゆみ, 株式会社ホットリンク 榊剛志

発表要旨

本公演では、ブログの書き込み頻度時系列に数理構造—特に、ゆらぎと拡散—について物理学的視野でみた観測とモデル化、また、その応用について報告する。具体的には、論文[1]と[2]の2つの論文の成果の報告とできれば最近えられた時系列の拡散特性に関する新しい成果について報告したい。

[1]はブログのゆらぎに関する数理基礎解析の結果、[2]はその成果等のマーケティング応用、具体的には、食の流行解析への応用である。

以下、本発表で扱う論文[1]と[2]のそれぞれについてその要旨を記述する。

(1) ブログのゆらぎと拡散の基礎的な性質[1]

一国規模のブログの特定の単語の書き込み数の日次の時系列のランダム成分は、テラー則と呼ばれるべき乗則が観測されることが知られている。テラー則とはゆらぎの大きさ（時系列の標準偏差に対応）が着目システムのサイズ（時系列の平均に対応）のべき乗に比例するという法則である。テラー則は、川の流量やインターネットパケット流量など自然法則やソーシャルメディアデータなど自然現象社会現象問わず広く観測されている。これまでの研究により、テラー則は、ランダム拡散モデルと呼ばれるポアソン分布の混合モデルの一種で説明できることが知られていた。しかし、ランダム拡散モデルは、定常モデルであり、非定常なブログ時系列にそのまま適応することには問題があった。

そこで、論文（1）では、ブログ系でのテラー則を詳細に記述するためランダム拡散モデルを非定常に拡張を試み、そのモデルを数理解析と過去9年間の日本語ブログ30億記事中での2000語の形容詞の時系列で検証した。はじめに、簡単なブロガーの行動モデル（マイクロモデル）から導出出来るランダム拡散モデルの非定常拡張を導入した。次に、拡張モデルをデータに適応することで、（1）各単語のゆらぎは平均書き込み数という一つのスケールパラメータのみで基本的には特徴づけられること、（2）モデルは、既知のブログにおけるいくつかのゆらぎのスケール特性をすべて同時に説明できること、（3）モデルを用いることで、これまで知られていなかったブログ時系列の特性を見いだせること、を示した。ここで、これまで知られていなかった特性とは、例えば、単語の平均書き込み数が大きい場合、従う分布は単語の種類に依存せずt分布に近い普遍的な分布に漸近すること等である。

加えて、論文では、上記基礎数理特性研究の応用として、2000語の形容詞に対して、特別なイベントがないと仮定したモデルから導出できるゆらぎの大きさと現実との観測の乖離を測定することによ

る国民的イベントのインパクトの計量方法も提案している(図 1)。

講演では、様々な分野の集まる学会の特性を考え、数理的詳細より、応用や現象を中心に紹介したい。

(2) ブログ時系列解析の食の流行解析への応用[2]

近年、企業や様々な組織において、公式アカウント等を利用した広報・コミュニケーション活動など、ソーシャルメディアのマーケティング活動での利用はもはや当たり前になりつつある。ソーシャルメディアのマーケティングでの利用法の一つにソーシャルリスニングがある。ソーシャルリスニングとは、ソーシャルメディアデータ(SNS データ)を用いて社会や生活者・消費者の“思い”を調査する活動である。企業活動においては、消費者の生に近い声やリアルタイムの反応を知れる等の長所があるため、広告の効果測定、製品の感想調査、経営リスク発見等様々な用途に利用されている。

一方、社会の期待に十分応えられていないソーシャルリスニングに関するニーズも存在する。その一つが流行把握と予測のニーズである。流行把握と予測のニーズとは、「世間全体での流行の現状や今後の行方を大規模データを用いて、これまでの手法より、客観的かつ精密に即時的に知りたい」というニーズである。このニーズは、例えば新商品開発部門においては「流行の兆しを 6 か月前までに知ることにより、商品開発の生産性をあげたい」というより具体的なニーズに結び付いてくる。

これまでもこのようなニーズや解決法も存在したが、その高度化の手段として SNS データが期待される理由は以下の特徴からである：(1) 分野やジャンルに依存しない広い情報を含んでいる[広く社会全体の流行の全体像を知れる](2) 売上データ等の結果のデータでなく、動機や思いの情報を含んでおり、それを定量化できる可能性がある[流行に至る理由やその背景を知ることができる](3) 半リアルタイムに情報を収集できる[今の状態を知れる]。

そこで発表では、論文[2]で議論している SNS データによる流行把握予測ニーズに応える第一歩のプロトタイプとしての「食の流行や関心の現状分析システム・サービス」について報告を行う[図 2, 図 3]。具体的には、特に、「日本中のブログに書かれる食べ物に関する語から、人知れず、長期的に関心が増加し続けている食に関する言葉を探す」技術—1. 食に関するキーワード辞書生成技術(新し食語の発見タスク), 2. 食に関するキーワードの件数時系列なから徐々に件数が増加している語[関心が増えていく語]を流行語候補として抽出・順序づけして出力する技術(絞り込み・流行候補語ランキングタスク)—を中心に議論したい。

参考文献

[1] Hayafumi Watanabe, Yukie Sano, Hideki Takayasu and Misako Takayasu, “Statistical properties of fluctuations of time series representing the appearance of words in nationwide blog data and their applications: An example of observations and the modelling of fluctuation scalings of nonstationary” ax time series, Physical Review E 94, 052317 (2016)

[2] 渡邊隼史, 小森あゆみ, 榊剛志, 大規模ブログデータベースを用いた食の流行の現状把握システムの開発, 日本マーケティング学会 カンファレンス・プロシーディングス vol.5 (2016) ,pp.162-174

図1（出所論文[1]）：ブログ時系列ゆらぎと国民的イベント。

上図．形容詞2000語の使用の“異常度”を測定による国民的イベントの定量化。図は横軸が日付で縦軸が異常度。上にいくほど異常である。例えば、「赤四角」の3月11日の異常度をみると、2011年で急増し、その後、通常の日にはちの特性に緩和しつつあることが確認できる。具体的には、2011年に「青三角（白抜き）」で示した大晦日程度だったものが、2014年には、「緑三角（塗りつぶし）」で示したクリスマスと「紫十字」で示した母の日の間程度のレベルに下がっている。また、ほかにも、「緑ひし形」で示したハロウィーン（10月31日）が2010年以降、異常度が増し、2014年では母の日程度になっていることが確認できる。下図：形容詞2000語のアンサンブルスケーリングにみる東日本大震災直後の一か月。地震発生の日から時間がたつにつれて平常時の理論線（桃色破線）に緩和していることが確認できる[黒三角（2011年3月11日）、赤丸（3月16日）、青四角（3月29日）]。挿入小図は理論との乖離度の時間依存性（詳細は論文[1]参照）を示す。この図からも乖離が時間共に縮小していることが確認できる。

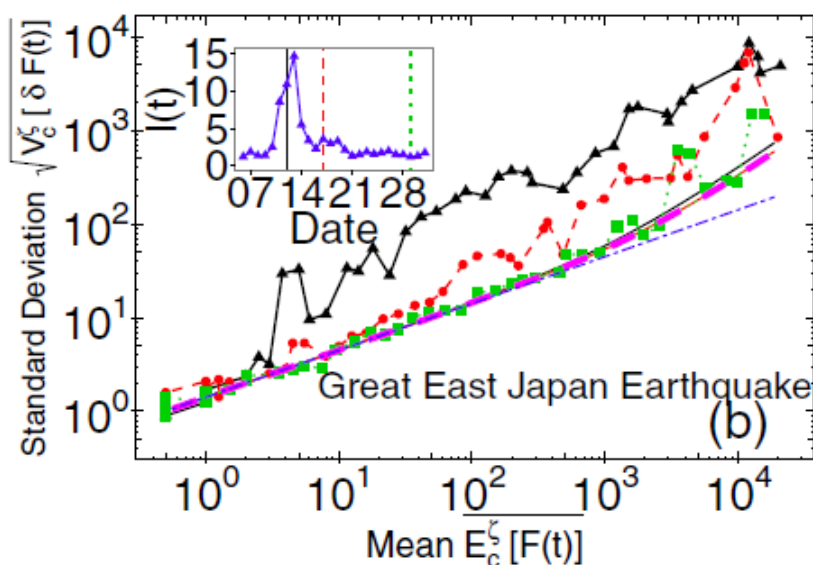
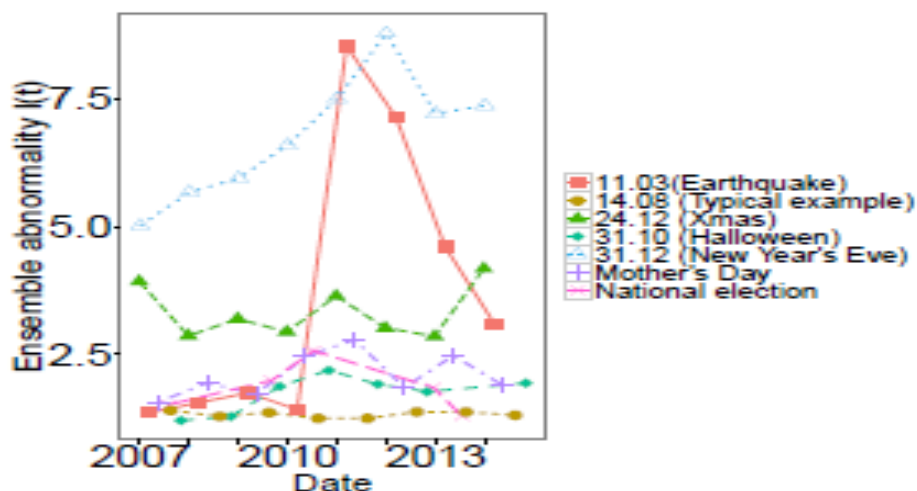
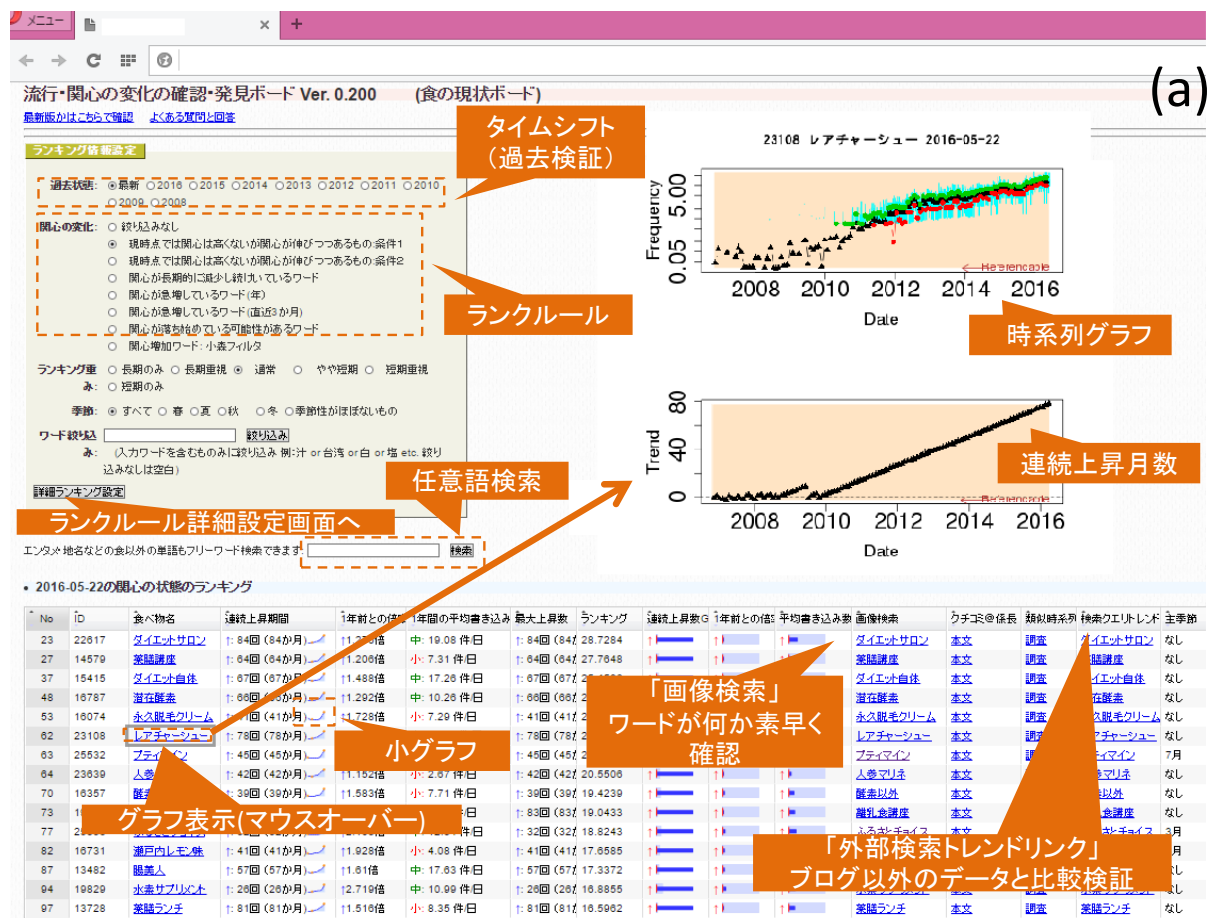


図 2 (出所論文[2])： 流行把握・表示システムとその応用。

(a) 食の現状ボード Web (b)流行解析システムのシステム構成 (c)流行予兆候補レポートの例



(b) 食の流行の現状観測システム

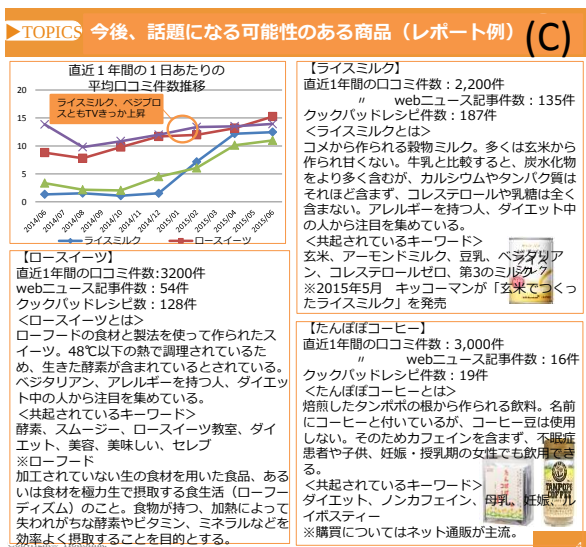
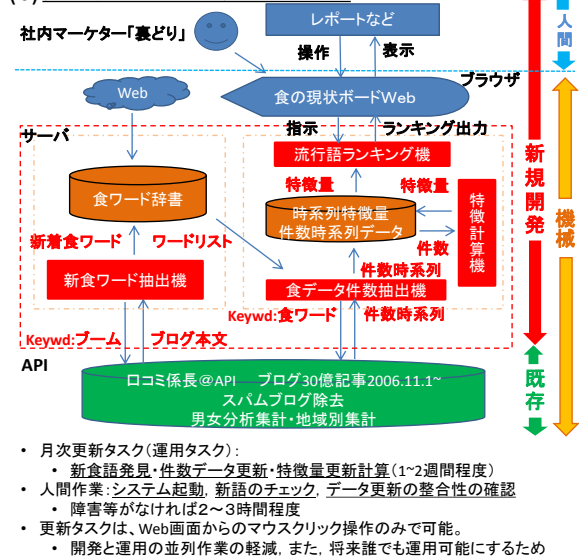


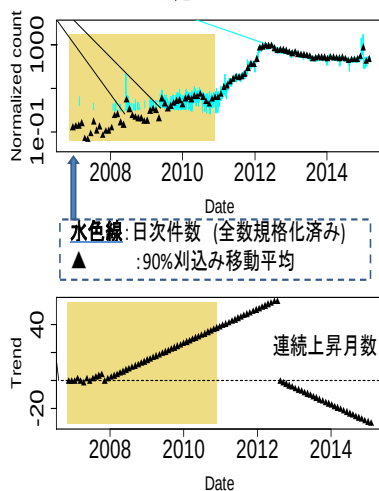
図3 (出所論文[2]): 過去データをもちいた流行の早期発見の検証。

(a)2010年11月時点での件数が大きく継続的に書き込みを増やしている語(下線は着目時点の2010年以降1年伸び続けた語。「最大上昇月数」は2015年11月までに到達した最大連続上昇月数)。(b)塩麹と(c)糖質制限の時系列(黄土色の四角つなかけ部分が使用したデータ範囲)。

(a)

順位	ワード名	連続上昇月数 2010	年増加率 2010	年平均件数 2010	最続上昇月数 2015	順位	ワード名	連続上昇月数 2010	年増加率 2010	年平均件数 2010	最大上昇月数 2015	順位	ワード名	連続上昇月数 2010	年増加率 2010	年平均件数 2010	最続上昇月数 2015
1	オムツケーキ	48	3.25	16.4	82	21	川越シェフ	18	16.3	17.3	26	41	お家ごはん	43	1.32	12.6	68
2	ケーキサレ	41	4.97	19.4	47	22	焼き小籠包	29	4.75	7.3	45	42	とんてき	49	1.25	5.56	65
3	バイセン	48	2.21	7.4	96	23	塩麹	36	1.95	4.9	56	43	まぜそば	46	1.23	14.3	103
4	川越達也	39	7.71	14.3	47	24	アヒージョ	48	1.80	6	105	44	牛骨ラーメン	36	3.68	2.37	64
5	赤飯さん	31	3.50	19.7	34	25	バーニャカ ウダソース	45	1.67	7.8	49	45	クリームチー ク	31	1.62	5.57	40
6	旦那さん弁当	49	2.11	6.96	77	26	マカロニタワ ニ	49	1.61	5.82	106	46	玉ねぎスープ	47	1.58	6.38	53
7	変形フレンチ	47	2.02	12.0	78	27	アイシング クッキー	41	1.89	11.9	98	47	ブレメル	34	2.42	5.51	81
8	かりんとうまん じゅう	38	3.94	6.00	42	28	酵素ドリンク	28	4.38	9	68	48	チーズトッポ キ	48	2.60	1.50	61
9	SABON	35	1.71	19.4	53	29	料理男子	29	2.87	2.1	51	49	米粉 100%	42	1.27	3.96	43
10	塚田農場	44	3.33	4.4	82	30	ホワイトフレ ンチ	49	1.71	4.9	86	50	GOPAN	12	53.1 4	18.2 5	16
11	かりんとう饅 頭	30	4.23	8.0	43	31	メガボテト	47	1.67	4	48	51	酵素ジュース	31	2.05	4.43	72
12	はま寿司	47	2.58	6.8	94	32	バラ焼き	45	1.44	6	49	52	酵素液	26	4.42	5.95	55
13	糖質制限	49	1.37	11.2	104	33	A5ランク	49	1.41	11	87	53	大豆ミート	40	1.71	3.13	97
14	アジנג	49	1.28	17.6	86	34	鰹掛け	49	1.26	9	97	54	吉田類	43	1.64	5.68	63
15	グリーンスム ニージ	38	5.19	12.9	61	35	トンテキ	49	1.23	15.0	70	55	丸源	47	1.53	13.3 2	63
16	着井	33	6.36	4.5	88	36	アガベシロ ップ	31	2.07	2.84	66	56	純豆腐	48	1.45	15.5 1	59
17	炭酸バック	33	2.98	17.6	44	37	ダイバーケ ーキ	38	1.81	3	47	57	多肉さん	46	1.45	6.92	103
18	朝ラー	48	2.76	8.4	49	38	サムギョブ ル	26	1.74	0.13	34	58	すた井	46	1.33	13.7 3	51
19	濃厚つけ麺	49	1.33	4.0	61	39	おんどる	49	3.89	7.5	60	59	かりん糖	49	1.31	6.25	61
20	森崎友紀	26	22.2	3.7	37	40	塩唐揚げ	49	1.8	2.5	73	60	ビッグボーイ	49	1.8	18.6	53

(b)「塩麹」の時系列
塩麹 2010-11-17



(c)「糖質制限」の時系列
糖質制限 2010-11-17

