

大規模ブログデータベースにおける単語の出現頻度の解析

—ノイズ成分の数理構造解析から食の流行把握への応用まで—

渡邊隼史

社会データ構造化センター(情報システム研究機構) 特任助教

h.wata@ism.ac.jp
arXiv:1604.00762 (2016)

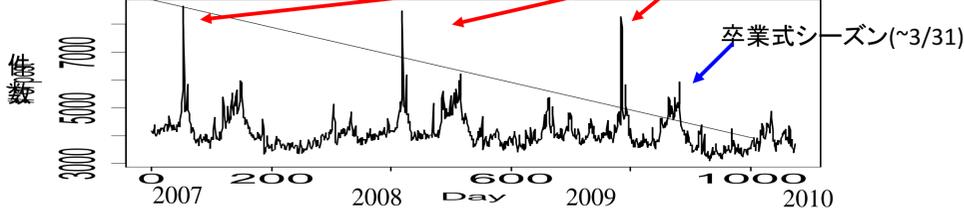
【ブログデータ】

- Web上の日記のようなもの。日付きのテキストデータ
- 2006年11月1日～現在までの日本語ブログ記事30億記事を利用
- データ収集は、ホットリンク社のロコミ係長を利用
- 基本的な量である着目キーワードの国内ブログ上での出現頻度時系列に着目

国内主要
ブログバイ
データを網羅

「さみしい」-感情の集計- 2007.11.1 ~

(全数規格化済み)



→集団としての人間の活動や感情を量的にとらえることができる。

【ブログデータの応用研究: 食の流行把握】 共同研究:小森、榎(ホットリンク)

食の流行の現状把握・予兆発見

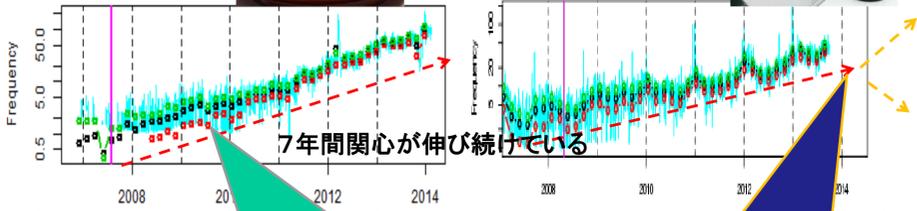
それをソーシャルメディアでできないか?

ニーズ (by 大手食品メーカー, by 雑誌流行)
消費者の動向の変化を他より早く知りたい
雑誌やTVにないような流行の兆しを知りたい

アヒージョ



獺祭(だっさい)



関心の増加を
ここで気づけたらいいな...

今のブームの過熱って
どのくらい? 未来は?

食の事項から、人知れず、関心が増えているものを探す
→食の辞書を作成し、件数が多くなく、長期間上昇し続けている語を抽出

「食」語を教える
(1) 単語の区切り
(2) 食べ物かどうか?

主に、以下の2つの技術を開発:

- 食べ物に関する新語候補発見技術:** 新しい食べ物語が次々登場への対応
「ブーム」と書かれたブログ記事に含まれる単語から新着の食べ物語を抽出
複合語処理(専門用語抽出)→「食ワード」と「非食ワード」の分類(LSA)
(1)味噌カレー牛乳ラーメン、オーガニックエクストラバージンココナッツオイル、昼飲み
(2)裏難波、キャロラインリーパー(唐辛子)、ファスティング(断食)
- 時系列連続的な関心増加(減少)の検出: 多様な時系列にロバストに統一的な処理の必要
長期的な単語の書き込み件数の上昇・下降の把握
瞬間的増加率でなく伸びの継続, 急増語では予兆としては遅い。

2010年11月の関心連続増加ランキング(件数 20件以下)

その後関心が1年以上伸び続けた語

オムツケーキ	48	3.25	16.4	82
ケーキサレ	41	4.97	19.3	47
パイセン	48	2.20	7.37	96
川越達也	39	7.70	14.2	47
赤飯さん	31	3.49	19.6	34
旦那さん弁当	49	2.10	6.96	77
変形フレンチ	47	2.02	11.9	78
かりんとうまんじゅう	38	3.93	5.94	42
SABON	35	1.71	19.3	53
塚田農場	44	3.32	4.43	82
かりんとう饅頭	30	4.23	8.01	43
はま寿司	47	2.57	6.79	94
糖質制限	49	1.36	11.1	104
アジング	49	1.28	17.5	86
グリーンスムージー	38	5.18	12.8	61
着井	33	6.35	4.54	88
炭酸パック	33	2.98	17.61	44

朝ラー	29	4.14	1.25	45
濃厚つけ麺	36	1.95	4.91	56
森崎友紀	48	1.79	6.06	105
川越シェフ	45	1.66	7.80	49
焼き小籠包	49	1.61	5.82	106
塩麩	41	1.89	11.9	98
アヒージョ	28	4.37	9.16	68
バーニャカウダソース	29	2.87	2.06	51
マカロンタワー	49	1.70	4.87	86
アイシングクッキー	47	1.67	3.54	48
酵素ドリンク	45	1.12	5.51	10
料理男子	41	1.89	11.9	98
ホワイトフレンチ	49	1.67	3.54	48
メガポテト	45	1.12	5.51	10
バラ焼き	41	1.89	11.9	98
A5ランク	49	1.67	3.54	48
館掛け	41	1.89	11.9	98
ドンテキ	49	1.67	3.54	48
アガベシロップ	49	1.67	3.54	48
ダイパーケーキ	49	1.67	3.54	48
サムギョプサル	49	1.67	3.54	48
おందор	49	1.67	3.54	48
塩唐揚げ	49	1.67	3.54	48
お家ごはん	49	1.67	3.54	48
とんてき	49	1.67	3.54	48
ませそば	49	1.67	3.54	48
牛骨ラーメン	49	1.67	3.54	48
クリームチーズ	49	1.67	3.54	48
玉ねぎスープ	49	1.67	3.54	48
プレモル	49	1.67	3.54	48
チーズトッポギ	49	1.67	3.54	48
米粉100%	49	1.67	3.54	48

長く伸びてる語
は次も上昇しやすい?

一期先上昇確率

連続上昇月数

連続上昇月数

連続上昇月数

連続上昇月数

連続上昇月数

連続上昇月数

連続上昇月数

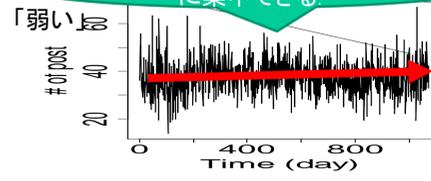
連続上昇月数

連続上昇月数

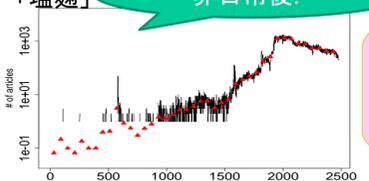
【ブログデータの基礎研究: なにもイベントない語のゆらぎの性質】

共同研究: 佐野(筑波)、高安(ソニーCSL)、高安(東工大)

特別なトレンドがなくゆらぎの性質
に集中できる



「塩麩」 非日常後

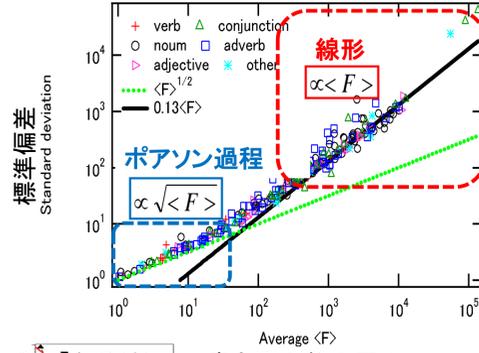


特別をよく
知るため、
まず通常を
知る

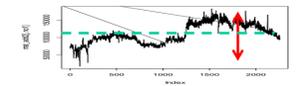
ーブログにおけるゆらぎの(時間)スケーリングとは:

- 平均書き込み数とゆらぎ(標準偏差)の間のべき乗関係 (ゆらぎのスケーリング FS)
- (1) 平均が小さい領域では0.5乗(ポアソンの) $\sigma_F \propto \sqrt{\mu_F}$
- (2) 平均が大きときは比例 (非ポアソンの) $\sigma_F \propto \mu_F$

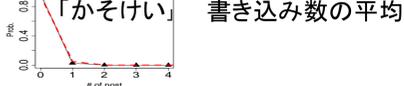
あるキーワード
が1日当たり日本中
のブログに
何回かかれるか?



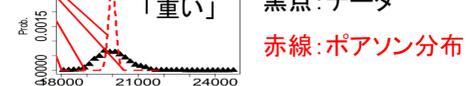
様々な言葉の書き込み数の
平均書き込み数 μ_F
とゆらぎの大きさ
の関係を調べた σ_F



非ポアソンの



珍しい単語



頻出する単語(流量の大きい川に対応)

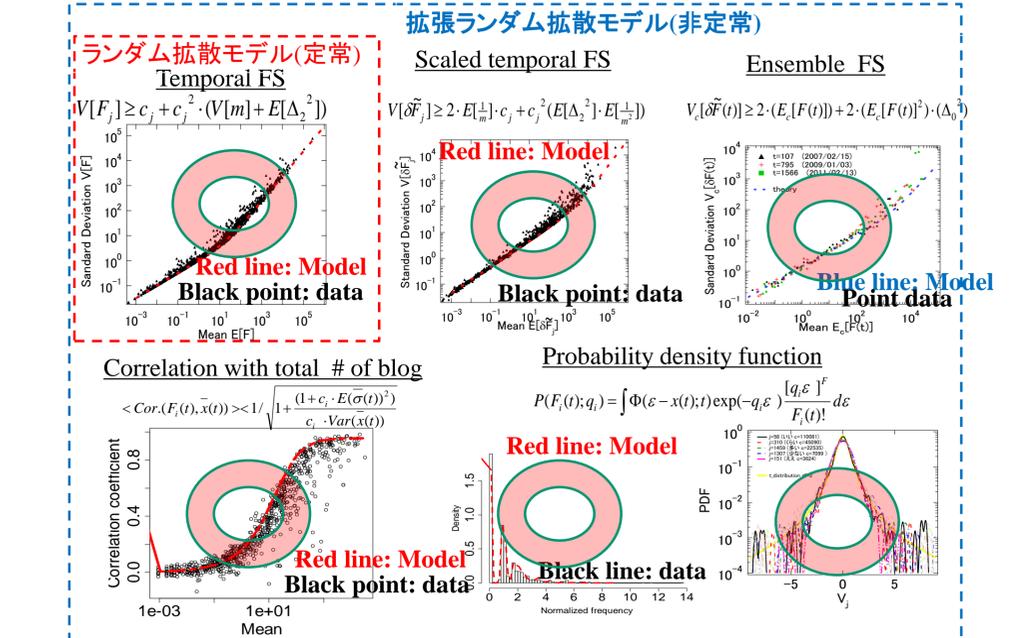
【ブログデータのゆらぎの性質: 拡張ランダム拡散モデル】

$$P(F_i(t); c_i) = \int \Phi(\varepsilon - x(t); t) \exp(-\varepsilon c_i) \frac{[c_i \varepsilon]^F}{F_i(t)!} d\varepsilon$$

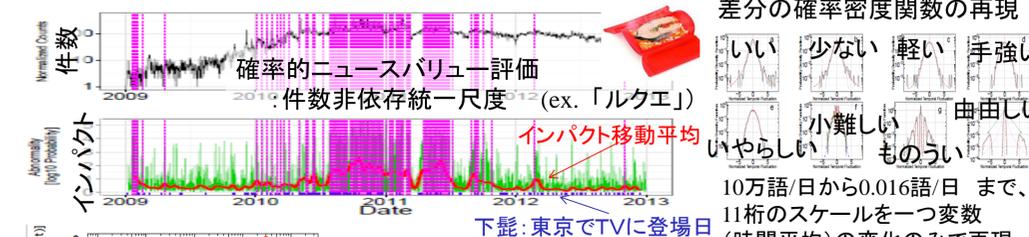
ゆらぎのスケーリングを説明するモデル [S.Meloni et al PRL2008, Y. Sano et al 2012 JEIC]
川の流量、インターネットパケットetc、複雑ネットワーク上での輸送ゆらぎの平均場近似

- ポアソン分布の重ね合わせ
- 個々のブロガーがランダムにポアソン分布で書き込む効果
- と社会全体のブロガー数がゆらぐ効果の、
- 重さ重ね合わせとして、書き込み数の分布決定

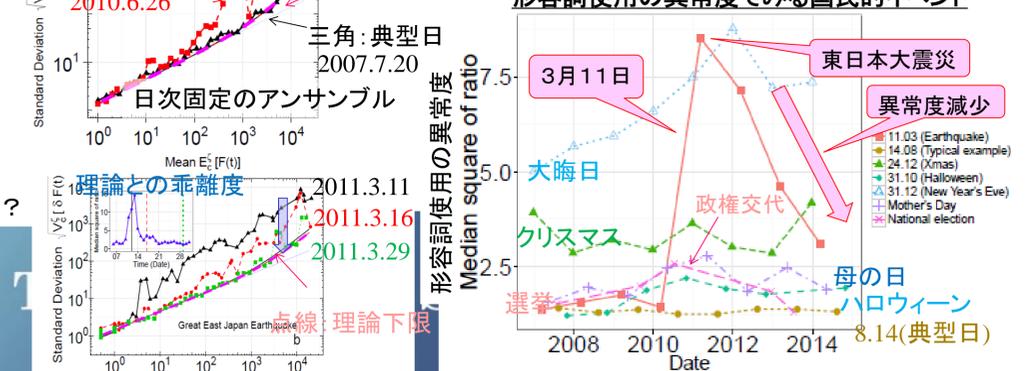
$F_i(t)$: 単語i書き込み数
 $c_j(t)$: 単語jの書き込み数平均
 $x(t)$: 規格化システムサイズ
 $\varepsilon(t)$: システムサイズの乱数成分



【ブログデータのゆらぎの性質と拡張ランダム拡散モデルの応用】



形容詞使用の異常度でみる国民的イベント



Application2: Systematic detection of vogues of foods

-We apply RD model to the noise reduction and trends detections-

