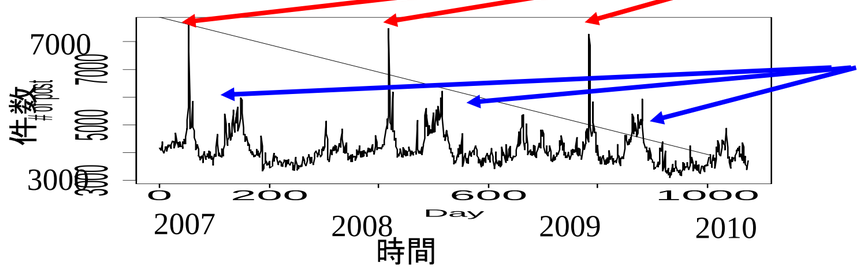
一国規模のブログデータからみる形容詞の使用頻度時系列データの拡散特性解析
～十分に日本語として定着した単語は一日一日どのくらいづつ使われ方が変化しているか？～

渡邊隼史　リスク解析戦略研究センター（統数研），社会データ構造化センター（情・シス機構）　特任助教
h-wata@ism.ac.jp

【ブログデータ】

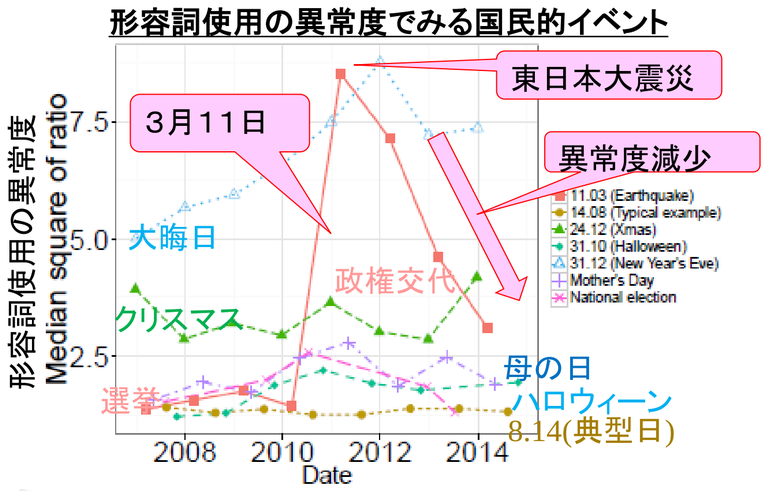
- Web上の日記のようなもの。 日付付きのテキストデータ
- データ収集は、ホットリンク社の口コミ係長を利用
- 2006年11月1日～までの日本語ブログ記事30億記事
- 基本的な量である着目キーワードの国内ブログ上での出現頻度時系列に着目

「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。

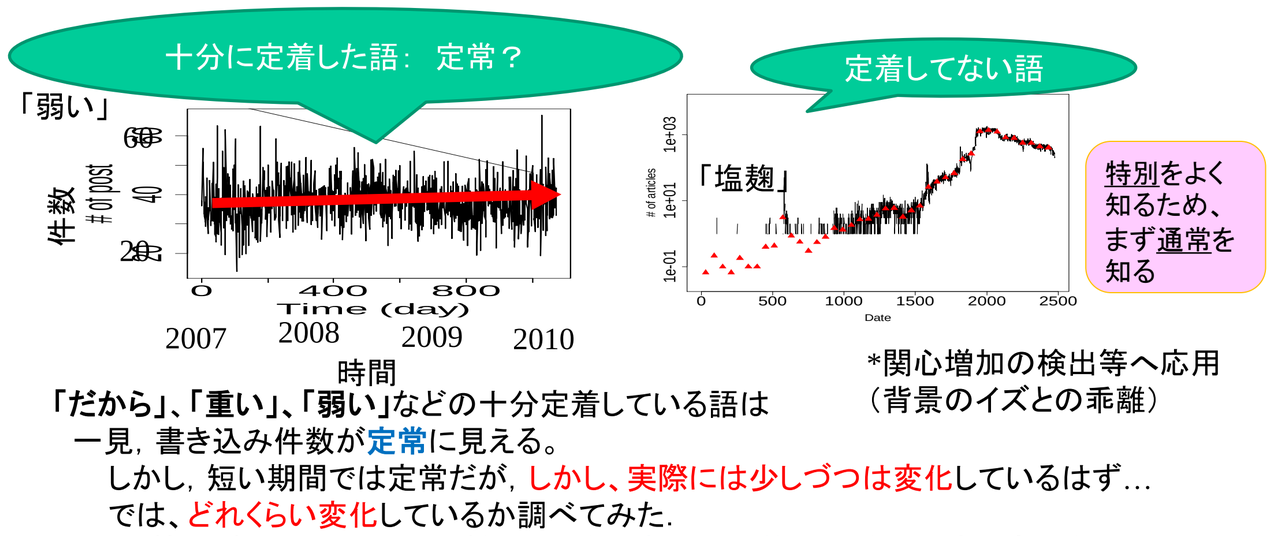
【日本語使用の異常度と国民的イベント―震災の風化の定量化―[1]】



＊「風化」への対応急ごう
福島民報
2015/03/15

＊日本人の秋祭り好きにはまった
ハロウィン バレンタイン越え
ダイヤモンド・オンライン
2014/10/30

【ブログデータの基礎研究：－十分定着した語の拡散特性－】



*関心増加の検出等へ応用
（背景のイズとの乖離）

「だから」、「重い」、「弱い」などの十分定着している語は
一見，書き込み件数が**定常**に見える。
しかし，短い期間では定常だが，しかし、実際には少しづつは変化しているはず…
では、どれくらい変化しているか調べてみた.

- 特別が外力やニュースがないような理想的な状態での単語の時間変化を知りたい。

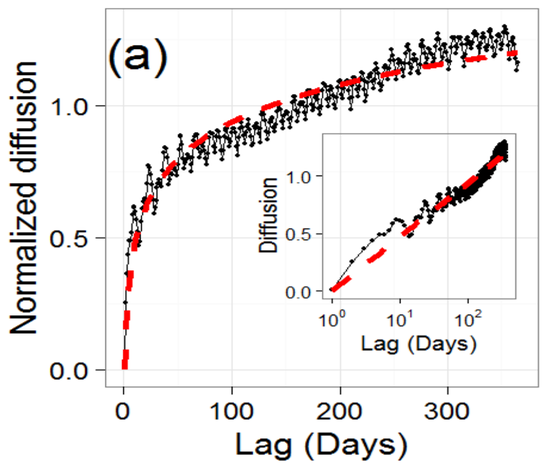
【研究の結論－対数拡散－】

- 拡散特性の観測：実際に，**着目日**の**L日語**にどのくらい書き込み数の平均的変化を観測

$$\sigma^2(L) = <(F_j(t+L) - F_j(t))^2>$$

t+L日の書き込み数が
ある時間tの書き込み数に対して
どれくらい変化しているか？
すべての単語は個々特別の事情があるがたくさんの単語集めればそれは打ち消す

（たくさんの種類の十分定着した語を測定し平均すると）



定常ではないが
とてもゆっくりした拡散

対数的に変化

予想される背後の数理構造は？

$$<(F_j(t+L) - F_j(t))^2> \propto \log(L)$$

【ブログデータの応用研究： 食の流行把握・予兆発見[2]】

食の事項から、人知れず、関心が増えてるものを探す（**レポーティングサービス**）
（1）**機械**：食の辞書を作成し，件数が多くなく，長期間上昇し続けてる語を抽出
（2）人間：レポート作成：裏どり，理由把握（今度の先行きの予想）など

理想は、流行確率〇〇%という魔法の方程式だが、魔法は使えない…

- 主に、以下の2つの技術を開発：

1. 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）
2. 時系列**継続的な関心増加（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.

「食」語を教える
（1）単語の区切り
（2）食かどうか？

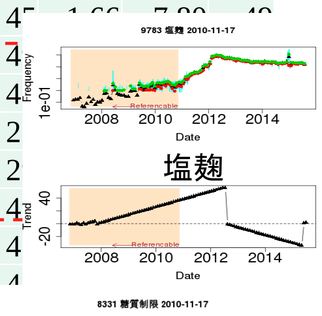
2010年11月の関心連続増加ランキング(件数 20件以下)
その後関心が1年以上伸び続けた語

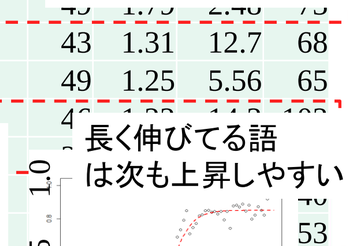
オムツケーキ	48	3.25	16.4	82
ケークサレ	41	4.97	19.3	47
パイセン	48	2.20	7.37	96
川越達也	39	7.70	14.2	47
赤飯さん	31	3.49	19.6	34
旦那さん弁当	49	2.10	6.96	77
変形フレンチ	47	2.02	11.9	78
かりんとうまんじゅう	38	3.93	5.94	42
SABON	35	1.71	19.3	53
塚田農場	44	3.32	4.43	82
かりんとう饅頭	30	4.23	8.01	43
はま寿司	47	2.57	6.79	94
糖質制限	49	1.36	11.1	104
アジング	49	1.28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

列：食べ物名，連続上昇月数，年増加率，年平均件数(2010.11時点)，歴代最大連続上昇月数(2015.11時点)

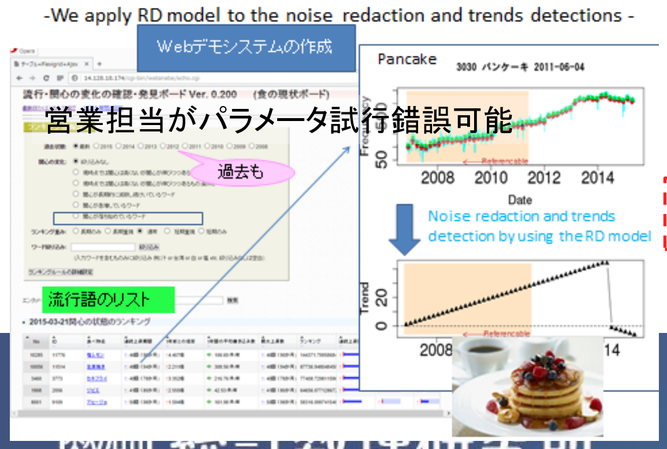
レポートの例

朝ラー				
濃厚つけ麺				
森崎友紀				
川越シェフ				
焼き小籠包				
塩麹	36	1.95	4.91	56
アヒージョ	48	1.79	6.06	105
バーニャカウダソース	4[illegible]			
マカロンタワー	4			
アイシングクッキー	4			
酵素ドリンク	2			
料理男子	2			
ホワイトフレンチ	4			
メガポテト	4			
バラ焼き	4			
A5ランク				
餡掛け				
トンテキ				
アガベシロップ				
ダイパーケーキ				
サムギョプル				
おんどる				
塩唐揚げ				
お家ごはん	43	1.31	12.7	68
とんてき	49	1.25	5.56	65
まぜそば	4[illegible]			
牛骨ラーメン				
クリームチーク				
玉ねぎスープ				53
プレモル				81
チーズトッポギ				61
米粉100%				43





Application2: Systematic detection of vogues of foods
-We apply RD model to the noise redaction and trends detections -



【対数拡散の背後の数理構造とモデル化－長期記憶と半整数差分－】

ランダム拡散モデル（川の流量ゆらぎ)[1]
-ブロガー集団の時間変化

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

ポアソン分布の混合　-点過程
x(t) ブログ全数

対数拡散の時間発展（乱れた媒質での拡散）
-単語jへの関心（需要）の変化

$$\frac{d^{0.5}c_j(t)}{d^{0.5}t} = \eta(t)$$

F(t) 書き込み数
C(t) 平均書き込み数
η(t) 平均0ランダムノイズ
-1/2回差分（微分）方程式，長期記憶，FIMA(0.5)

$$\frac{d^{0.5}c_j(t)}{d^{0.5}t} = \eta(t) \Leftrightarrow c_j(t) \approx \sum_{s=0}^{\infty}(s+a_0)^{-\beta}\eta_j(t-s)$$

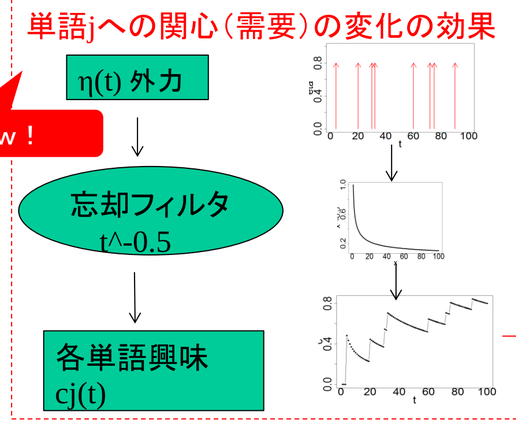
$$a_0 = \Gamma(\beta)^{-1/\beta}$$

$$\beta = 0.5$$

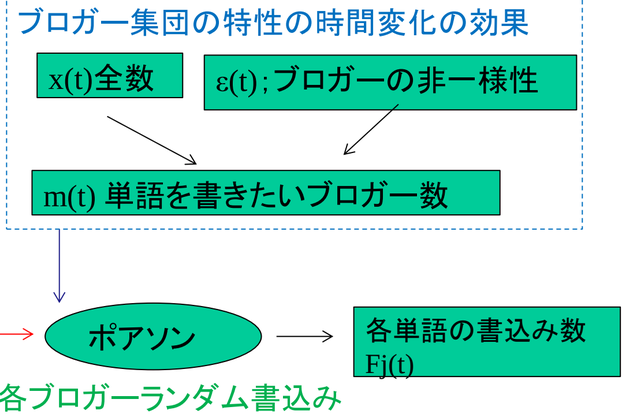
βのべき乗で外力を忘れていく確率過程

文献[1]

単語jへの関心（需要）の変化の効果

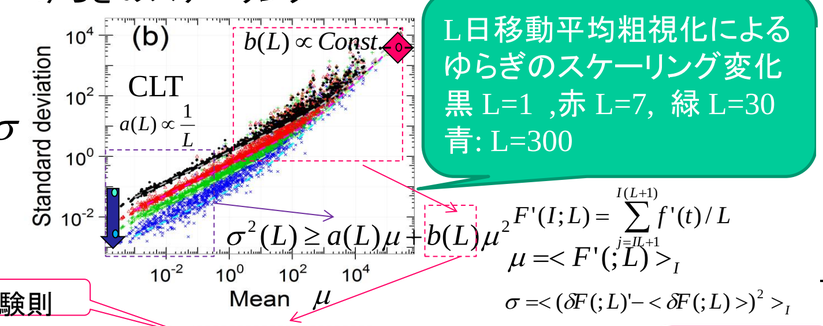


ブロガー集団の特性の時間変化の効果



【モデルの性質と実データの拡散以外の特性の再現】

－ゆらぎのスケーリング

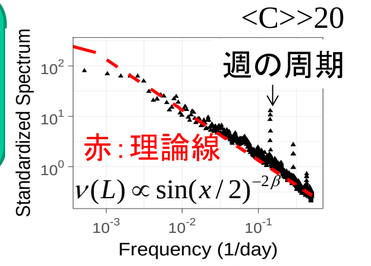


$$F'(I;L) = \sum_{j=lL+1}^{I(L+1)} f'(t)/L$$

$$\mu = <F'(;L)>_I$$

$$\sigma = <(\delta F(;L)' - <\delta F(;L)>)^2>_I$$

－（刈込）平均標準化スペクトル
<C>>20
週の周期



－F(x+L)-F(x)の密度分布