[1] Phys. Rev. E 94, 05231 (2016)　[2] 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）　arXiv:1801.07948 (2018)

2018年6月15日　統計数理研究所　オープンハウス

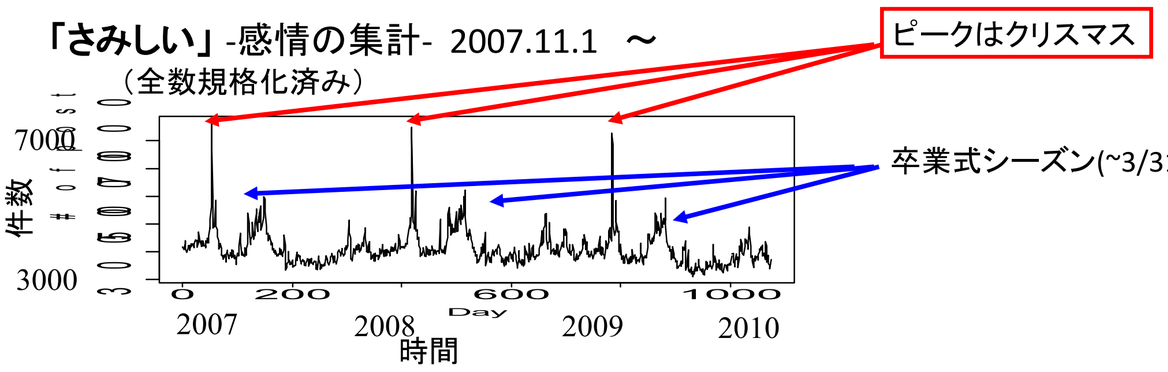
十分定着した単語は使われ方は日々どのくらいづつ変化しているか？
—様々な言語や媒体に共通してみられる単語出現頻度の対数拡散と非整数微積分—

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

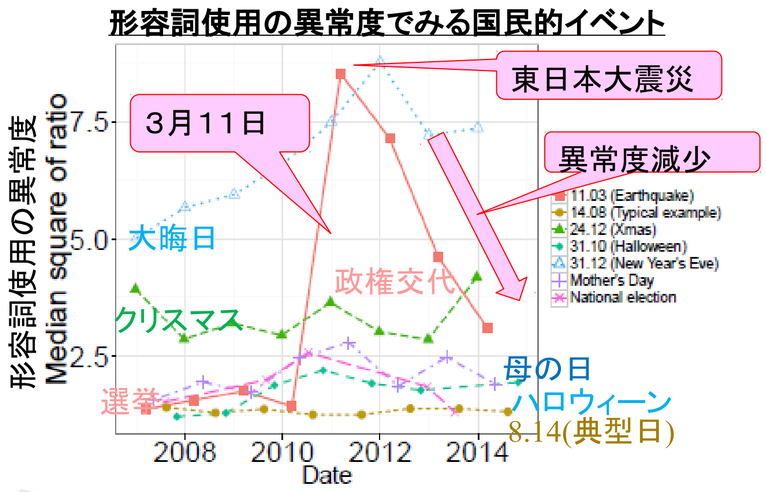
「さみしい」-感情の集計- 2007.11.1 ～
（全数規格化済み）



→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]】

形容詞使用の異常度でみる国民的イベント



＊「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

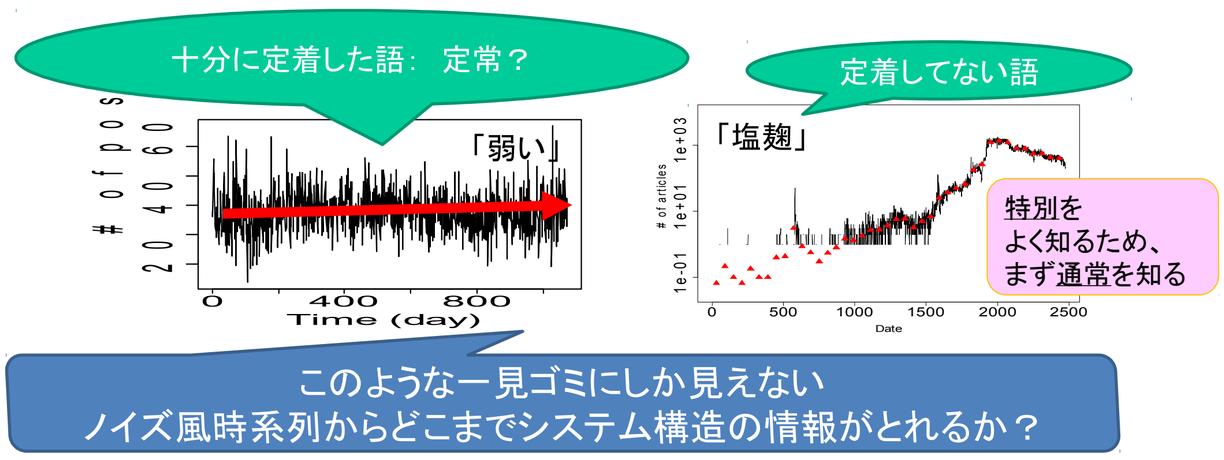
【研究の目的】

– 社会の安定性を定量的・数学的に調べたい.

- 社会は常に変化している部分もあるが、“コア部分“は安定しているように見える
- 例えば，「こんにちは」という挨拶はここ30年くらいはあまりかわっていない.
- では、**どのくらい安定してどのように安定してるかを定量的に調べてみたい**.
  – “変わらないや安定“は社会の基盤なので目立つ大きな動きと同様に社会の動的な性質の理解に重要と考えられる.

– 今研究では、そのひとつの例として，言葉において「十分定着した単語」はどのくらい安定しているかをデータを用いて精密に測定.

- **例えば，「だから」、「重い」**などの十分定着している語は
  日本のブログでの書き込み件数がほぼ定常で安定してるように見える。
  –しかし，少しづつは変化しているはずで…
  –実際にはどのくらい変化してるかを日次の解像度で調べてみたい.

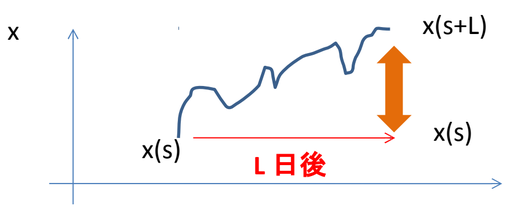


【手法: 平均二乗変異】

- 今回の研究では**「平均二乗変位」**という時系列量を測定した
  「平均二乗変位」は，**着目日のL日語**にどのくらい書き込み数が
  平均的に変化するかをあらわした量（**拡散特性**）

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

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？
平均



$$\sigma^2(L) \approx \sum_{t=1}^{T-L} \frac{(x(t+L)-x(t))^2}{T-L}$$

色々なLについて計算

- べき忘却ランダムウォーク（積分形）

$$r_j(t) = \frac{1}{\Gamma(1-\beta)} \sum_{s=0}^{\infty} (s+d_\beta)^{-\beta} \cdot \eta_j(t-s),$$

⇔

- 微分形

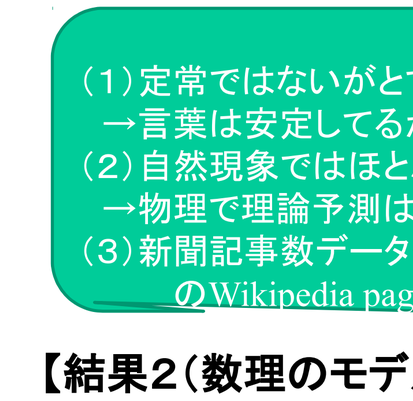
$$\frac{d^{1-\beta}}{dt^{1-\beta}} r(t) = \eta(t),$$

$$\left(\frac{d^{1/3}}{dt^{1/3}}\right)^3 x(t) = \frac{d^{1/3}}{dt^{1/3}} \frac{d^{1/3}}{dt^{1/3}} \frac{d^{1/3}}{dt^{1/3}} x(t) = \frac{d}{dt} x(t).$$

【結果1（データの解析）：十分定着した単語の拡散特性】

●時系列
●ランダムウォーク（熱拡散、為替）

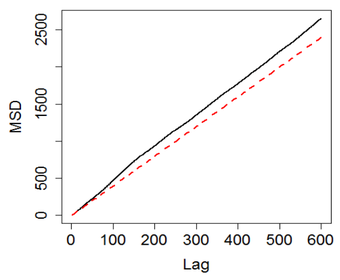
$$x(t+1) = x(t) + \varepsilon(t)$$



ε(t)：正規分布：平均0, 標準偏差2
x(1)=100　　：

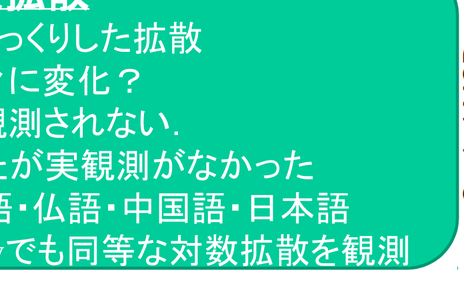
●**平均二乗変位**

$$\sigma^2(L) \propto L$$

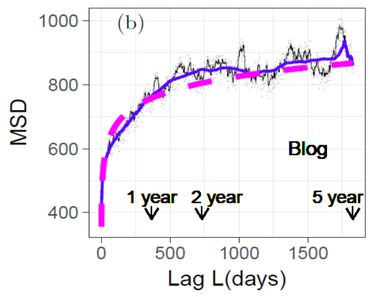


不安定

●**ブログデータ**
• **真田(key word)**

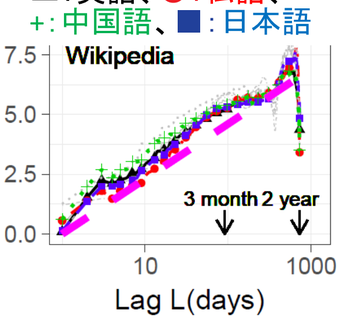


$$\sigma^2(L) \propto \log(L)$$



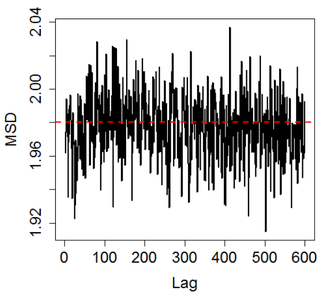
• 無相関定常ノイズ
（毎日独立にずっと同じ）

$$x(t) = \varepsilon(t)$$



ε(t)：正規分布：平均100
標準偏差2

$$\sigma^2(L) \propto O(1)$$



安定

対数拡散

（1）定常ではないがとてもゆっくりした拡散
→言葉は安定してるが徐々に変化？
（2）自然現象ではほとんど観測されない.
→物理で理論予測はあったが実観測がなかった
（3）新聞記事数データや英語・仏語・中国語・日本語のWikipedia page viewでも同等な対数拡散を観測

▲：英語、●：仏語、
+：中国語、■：日本語

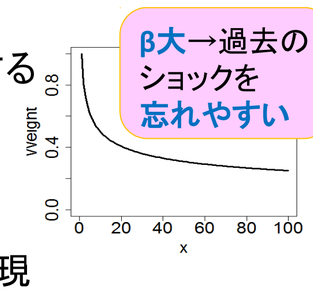
【結果2（数理のモデル解析）：非整数微分との関係】

• **べき忘却ランダムウォークモデル**が性質をよく再現する

$$r_j(t) = r_j(0) + \sum_{s=1}^{t} (t-s+1)^{-\beta} \eta_j(s)$$

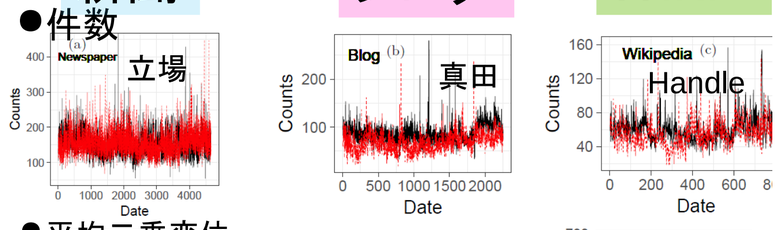
→とくにブログのような対数拡散は**忘却速度β=0.5**

*時系列は「べき忘却＋点過程（RDモデル）」でよく再現

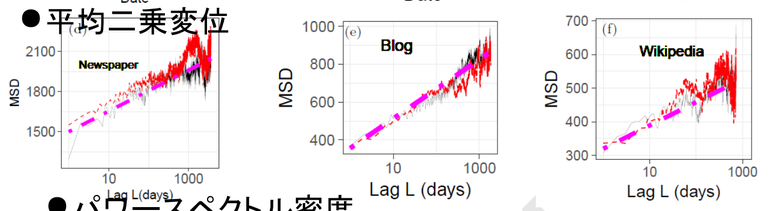


新聞	ブログ	Wiki

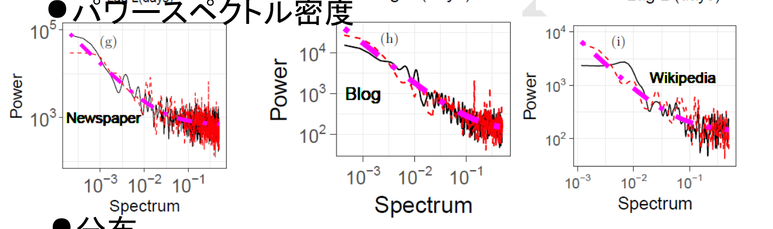
●件数



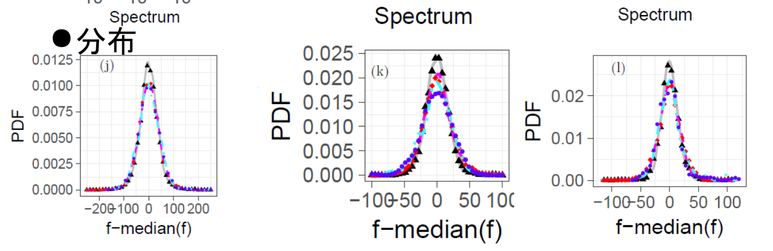
●平均二乗変位



●パワースペクトル密度



●分布



導入したランダムウォークモデルで本質的には，

新聞、ブログ、Wikipedia(4言語)の**二乗変位**，**スペクトル**、**分布**という統計特性が**統一的に**再現できることがわかった.

媒体、単語、言語等の相違は**2つのパラメータ**で表現できる.

黒実線：データ
赤点線：シミレーション
桃点破線：理論

＊シミレーションでは点過程（ランダム拡散モデル）のrateとし忘却RWモデルを使った.

【考察：忘却速度β=0.5の意味】

（1）定常と非定常の境界（非常に微妙な立ち位置のパラメータ）

$$< (r_j(t+L) - r_j(t))^2 > \propto \begin{cases} L^{1-2\beta} & (0 \le \beta < 0.5) \\ \log(L) & (\beta = 0.5) \\ O(1) & (\beta > 0.5) \end{cases}$$

非定常
単語カウント
定常

忘れ方がβ=0.5より少しでも大きければ**「定常」**に
忘れ方がβ=0.5より少しでも小さければ**「非定常」**になる.

（2）ランダムウォークとIIDノイズのちょうど中間（半階微分）

• ランダムウォーク（1階微分）	• ブログなど（1/2階微分）	• IIDノイズ（0階微分）
$\frac{d}{dt} r_j(t) = \eta(t)$	$\sqrt{\frac{d}{dt}} r(t) \approx \eta(t),$	$\frac{d^0}{dt^0} r_j(t) = r_j(t) = \eta(t)$