様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年 5 月 25 日現在

機関番号：１２６０８
研究種目：挑戦的萌芽研究
研究期間：2010　～　2011
課題番号：２２６５６０２５
研究課題名（和文）インターネットの書き込みにおける人間集団行動の統計物理的考察
研究課題名（英文）　Statistical physics analysis of collective human behaviors observed in the internet information

研究代表者
高安　美佐子（TAKAYASU　MISAKO）
東京工業大学・大学院総合理工学研究科・准教授
研究者番号：20296776

研究成果の概要（和文）：ブログやツイッターなどの公開されているインターネットの書き込み情報を自動的に収集蓄積した巨大なデータを、統計物理学的な視点に基づき科学的な分析をした。まず、日常的に一定の頻度で使われていると期待される単語に注目し、その基本的なゆらぎの特性を明らかにした。次に、流行語やニュース用語などに注目し、その上昇・下降が指数関数やベキ乗の関数形で記述できることを示した。

研究成果の概要（英文）：From the viewpoint of statistical physics we analyzed the huge data of blogs and twitters. Firstly, we paid attention to ″daily words″ which are used in constant rates, and observed their statistical properties in the number fluctuation of appearance. Next, we focused on ″booming words″ and ″news words″ paying attention to the functional forms of growth and decay. We found that such growth and decay are described either by an exponential function or a power law.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,800,000	0	1,800,000
2011 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,000,000	360,000	3,360,000

研究分野：工学
科研費の分科・細目：応用物理学・工学基礎・工学基礎
キーワード：数理物理、ネットワーク、社会学、行動学

１．研究開始当初の背景

最近、インターネットのホームページに簡単に自由に個人的な見解を書きこめるブログという仕組みが誕生した。ユーザ数は急激に増加し、日本国内だけでも連日数百万人もの書き込みが行われるようになった。このような書き込みの内容を科学的に分析すれば、人間の集団的な行動が定量的に記述することができると考え、研究プランを立案した。

２．研究の目的

ブログの書き込み記事のデータ解析のための基礎的なデータ処理方法を確立する。次に、キーワードの出現頻度に関する経験的な法則を統計物理学の方法を用いて網羅的に探し出す。キーワードとしては、日常的に一定の割合で出現すると考えられる日常語、次第にブームになっていくブーム語、突然のニュースで急激に立ち上がるブーム語の３つのカテゴリーに注目し、それぞれのキーワー

ドの出現頻度の特性を解明し、それらの特性を再現するような数理モデルを開発する。

３．研究の方法

まず、データの非定常性を取り除き、システマティックな誤差を排除する基本的なデータ処理手法を確立する。ブログの書き込み数の算出には既存の検索エンジンを活用しているが、自動生成されたスパムを除外するフィルターを通し、さらに、サーバーダウンなどによるシステマティックな変動を補正する手法の開発が必要とされる。本研究では、書き込みされた全ブログ書き込み単語数で規格化することによって、システマティック誤差を補正する方法を導入した。

次に、基本的な単語出現頻度のゆらぎの特性を明らかにするため、出現頻度が最も安定していると期待される日常語に注目し、そのゆらぎの特性を明らかにする。

一定の頻度で出現する単語のゆらぎの特性が明らかになると、その特性から乖離した状態を確認することによって、非日常語を定義することができる。本研究では、ゆっくりと出現頻度が増加するような流行語や、突然、急激に出現頻度が高くなるニュース語に着目し、それらの増加や減少の特徴を抽出し、その関数形を推定する。

４．研究成果

スパムフィルターを通し、品詞分解した形でのブログの書き込みデータにおいてみられる典型的な単語の出現頻度の時系列を、図１に示す。

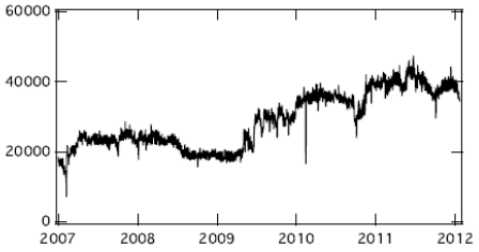

図１　ある単語の出現頻度（日次）の変化の例

この変動のグラフからは、この単語の出現頻度が次第に上昇しているように見える。しかし、同時期の全ブログ書き込み単語数の時系列は、図２に示されており、全体的な挙動が非常によく似ていることが確認できる。すなわち、図１における単語数の出現頻度の変動は、全単語数の変動に起因する成分が大きいことが推測される。

そこで、全単語数で規格化した単語の出現頻度に変換すると、図３のように、多少のノイズのようなパルスが残るものの、全数の変動に起因する個々の単語数の増減の成分を取り除くことができることがわかる。

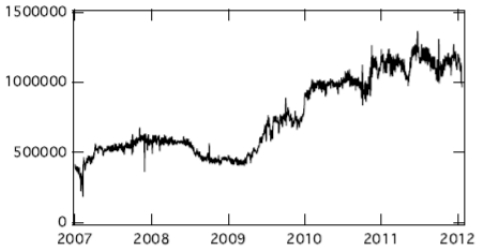

図２　全ブログに書きこまれた総単語数の日次での変動

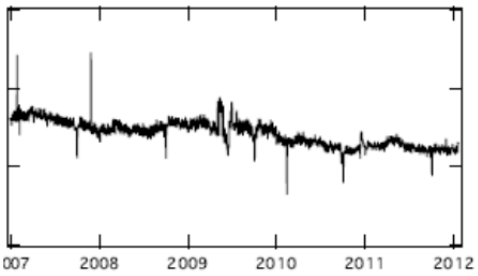

図３　図１の単語の出現頻度の変動を、図２の単語の総数で規格化した変動

次に、一定の頻度で出現すると期待される日常語に注目し、単語の出現頻度がどのように変動するのかを観測する。頻度の非常に低い単語から使用頻度のかなり高いものまでほぼ出現頻度が安定している単語を選び、上記の規格化を施した上で、その標準偏差を求め、単語の出現頻度の関数としてプロットしたのが、図４である。

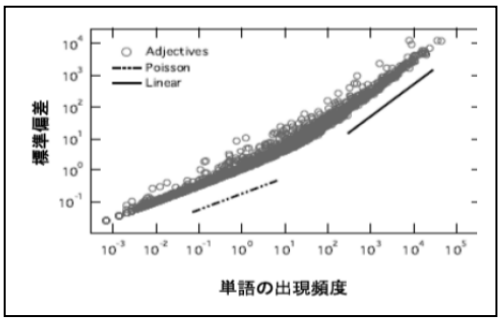

図４　一定の頻度で出現する単語の出現頻度とその数のゆらぎの標準偏差の比較

もしも、理想的な独立な確率的な事象であれば、書き込み数はポアソン過程にしたがうはずであり、標準偏差は出現頻度の平方根に比例することになる。図４から、このような関係が成り立つのは、出現頻度がある程度より小さい場合に限定され、それ以上の出現頻度の場合には、標準偏差が出現頻度に比例することがわかる。

このように出現頻度が大きい時に標準偏差が異常に大きくなる現象は、ランダム拡散モデルとよばれる数理モデルによって理解することができる。これは、ブログの書き込みをする可能性のある人の数自体がまず、ランダムに変動し、その中で書き込みをした人が特定の単語を使用するという２重のゆらぎが関与していることを意味する。モデルによれば、ブログの書き込みをする人の数の変動の割合から、図４の折れ曲がりの値が決まるので、逆に、観測データから、潜在的なブログの書き込みをする人の数の変動が推定できることになる。

次に、ブームによってゆっくりと出現頻度が増加するようなキーワードやニュースによって突然頻度が高くなった単語の人気がゆっくりと減衰する様子を観測した。

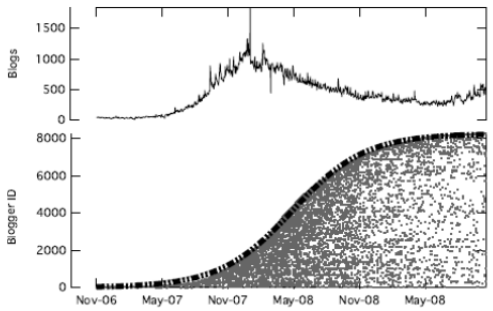

図５　ブーム語の増加と減衰の時系列の例。下段は、ひとりひとりの書き込みの様子をドットで表現している。

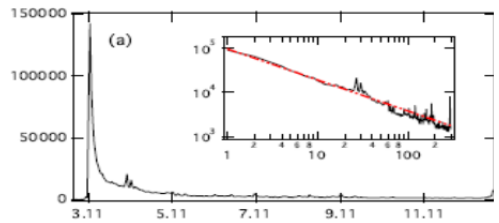

図６　ニュース語の急激な増加とその後のゆっくりした減衰。両対数プロットで直線的になり、時間のベキ乗で減衰していることがわかる。

図５は、１年程度の時間をかけてゆっくりと頻度が増加するようなブーム語の例であり、詳しく調べると、上昇部分と下降部分は指数関数によって、時間変動を近似できることが確かめられている。このような増加と減衰の動力学は、伝染病のモデルとしてよく知られている数理モデルに基づく数理モデルによって、再現することができる。

図６は、突然注目されたニュース語の場合の時間発展の例である。一日の間に劇的に増加した書き込み数は、その後、ベキ乗則にしたがって徐々に減衰し、一年後でも同じ法則性が持続していることがわかる。このようなベキ乗の減衰は、様々なキーワードに対して普遍的に観測される。さらに、　このようなベキ乗の減衰に関しても、伝染病モデルを改良した比較的シンプルな数理モデルによって再現することができる。

以上まとめると、ブログの書き込みという新しい研究題材に対し、データの観測方法という土台の部分を構築し、基本的な単語の出現頻度のゆらぎの特性を発見し、さらに、ブーム語やニュース語の特徴的な動力学的挙動をデータから見出し、それらを記述する数理モデルを考案することができた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
① Yukie Sano, Misako Takayasu
‘Macroscopic and microscopic statistical properties observed in blog entries’
Journal of Economic and Interaction and Coordination,　査　読　有　,Vol.5,Issue2, 2010,221-230

〔学会発表〕（計８）
①佐野幸恵、高安秀樹、高安美佐子
「Zipf 則と Heaps 則を用いた潜在語数の見積もり」
日本物理学会 2012 年年次大会
2012 年 3 月 26 日、関西学院大学
②高安美佐子
「大規模データから観測される人間社会の法則性」
電子情報通信学会
2012 年 3 月 21 日、岡山大学
③佐野幸恵、山田健太、渡辺隼史、高安秀樹、高安美佐子
「ブロゴスフィアにおける単語出現頻度のゆらぎ」
電子情報通信学会
2012 年 3 月 21 日、岡山大学
④山田健太、高安秀樹、高安美佐子

「大規模ブログデータから観測されるブームの形成とモデル化」
電子情報通信学会
2012年3月21日、岡山大学
⑤山田健太、佐野幸恵、高安秀樹、高安美佐子
「大規模ブログデータから観測される統計的性質とモデル化」
日本物理学会2011年秋季大会
2011年9月22日、富山大学
⑥佐野幸恵、高安秀樹、高安美佐子
「ツイッターを用いた単語ネットワークの時間発展」
日本物理学会2011年秋季大会、
2011年9月22日、富山大学
⑦Misako Takayasu
'Critical behavior observed in Collective Human Behavior'
21st Int'l Conference on Noise and Fluctuations, Toronto(招待講演)
2011年6月13日、Ryerson Univ., Toronto
⑧渡邊隼史、佐野幸恵、山田健太、高安秀樹、高安美佐子
「日本語ブログにおける日常語の書き込み頻度の統計性と Random Diffusion Model の拡張」
日本物理学会2010年秋季大会、
2010年9月24日、大阪府立大学

〔図書〕(計１件)
①Misako Takayasu
'Econophysics Approaches to Large-Scale Business Data and Financial Crisis'
Springer,2010,342

〔その他〕
ホームページ等
http://www.smp.dis.titech.ac.jp/intro.php

６．研究組織
(1)研究代表者
高安　美佐子 (TAKAYASU MISAKO)
東京工業大学・大学院総合理工学研究科・准教授
研究者番号：20296776
(2)研究分担者
なし
(3)連携研究者
なし