様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ８ 日現在

機関番号：１３３０１
研究種目：若手研究(B)
研究期間：2017～2020
課題番号：１７Ｋ１３８１５

研究課題名（和文）大規模ブログデータを用いた流行・普及現象の網羅的定量研究―新語時系列解析の応用―

研究課題名（英文）A quantitative study of diffusion processes of popular trends: an application of time series analysis of word counts in nationwide blog data

研究代表者

渡邊　隼史（Watanabe, Hayafumi）

金沢大学・電子情報通信学系・助教

研究者番号：３０７８３９５６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：ブログや新聞等の時間付きデータにおける単語の使用の時間変化について研究を行った。本研究の一番の成果は，十分定着した語の単語使用数が対数関数的な速度で変化していることを明らかにしたことにある。この法則が日本語ブログだけなく，国内新聞や英語はフランス語のWikipediaでも共通して観測されることも示した。さらに理論研究により，社会的記憶との関係性も示唆した。また，新語普及についても，簡単な微分方程式にで多くの時間変化が統一的に説明できることを示唆し，論文化に向けて研究を進めている。さらに，明治大正昭和の新聞データに関する共同研究を開始し，独自フォーマットデータからpdfデータへの変換を完了した。

研究成果の学術的意義や社会的意義
本研究の一番の主要成果である十分定着した語に関する対数関数的拡散現象の発見は，言語科学的には，言語の種類（日本語，フランス語，中国語，英語）や媒体（新聞，ブログ，Page view）に依存しない時間的な言語法則の新たな発見の可能性もある。また，物理学的には，理論的な研究はなれてきたが，現実の観測がほとんどなかったultra slow diffusion という現象の稀な現実観測例になっている。加えて，明治大正昭和の新聞データに関する言語OCR研究は，今後研究が順調に進み開発に成功すれば，社会学、言語学、情報科学など言語データを使う様々な分野で共通に使われる言語資源の提供につながる可能性もある。

研究成果の概要（英文）：　We studied the temporal change of word uses in time series data such as blogs and newspapers. The most important result of this study is that we found the logarithmic growth (i.e., ultraslow diffusion)　of the time-series of word counts of already popular words by analyzing three different nationwide language databases: (i) newspaper articles (Japanese), (ii) blog articles (Japanese), and (iii) page views of Wikipedia (English, French, Chinese, and Japanese). Through theoretical research, we also suggested a relationship between this observation and social memory. In addition, We have started a joint research on newspaper data from the Meiji, Taisho and Shouwa periods and transformed the original format data to pdf data.

研究分野：社会データ解析

キーワード：時間付きテキストデータ　大規模データ解析　ソーシャルメディア解析　時系列解析　拡散現象　複雑システム科学　言語データ　新聞データ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

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・コミュニュケーション活動など，ソーシャルメディアデータ(SNS データ)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。そのような SNS データを用いたまだ十分に実現できておらず研究も進んでいないマーケティング活用のニーズに「流行把握と予測」がある。「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。ＳＮＳデータのもつ

- 分野やジャンルに依存しない広い情報を含んでいること
- 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いること,
- 半リアルタイムに情報収集可能,

という，特徴を活用する。 申請者は、これまで，このような課題に対して科学的手法を提供することを目指し，①ブログにおける書き込み現象の基礎的な物理学的・統計的な基礎的性質の数理モデリング　②基礎研究結果等を応用した流行解析システムの開発という，基礎から応用につなげる接近法で研究を行ってきた。

２．研究の目的

大量蓄積された大規模データをもとに，世間全体での流行の現状や今後の行方をそのデータの解像度を活かし，既存の手法より，客観的かつ精密に即時的に知りたいというニーズ（流行把握ニーズ）が存在する。これまで申請者はこの問題に対して，30 億記事 7 年分の大規模ブログデータと自然言語及び時系列処理技術を用いた, 食に関する〝現状の流行物候補〟や〝予兆候補〟の発見手法の研究やそれを提示するシステム開発を行ってきた。本研究では〝候補提示〟から一歩進歩させ「その候補が現状どのような流行の過程にあり，今後どのようになっていくかを定量的にマーケターに示す」手法の実現を目的に研究を行う。特に，基礎となる新語の普及過程の物理的・統計的な性質を明らかにすることを中心の目的に研究を行った。

３．研究の方法

本研究は，過去 10 年の約 30 億の大規模ブログ記事データベースとその記事に含まれる単語の時間変化の網羅的なデータ解析及び数理モデル化を主な手法とする。さらに，その数理解析の結果を流行の解析に応用する。研究の状況の変化により，大規模ブログデータの他, 新聞の書き込み件数データ, 加えて, Wikipedia の各国語の page view data 研究対象に加えた。

４．研究成果

a)十分定着した語の時間変化の性質

”社会の安定的な構造”は，どの程度どのように安定しているのか？ この疑問の一例として，本研究では,「重い」や「赤い」や「私」のような”社会に十分定着した単語の使われ方の変化や安定性”について大規模データに基づく定量的な調査研究を行った。その結果，国内 9 割程度のブログデータ，日本の新聞記事データ，英語・フランス語・中国語・日本語の Wikipedia のページビュー（Wikipedia のキーワードのページが何回閲覧されたかの回数）において，多くの個別の単語や様々な単語の全体の変化の平均としての振る舞いとして，言語や媒体に依存せず”十分定着した語の使われ方”は対数関数的に変化（超慢拡散）することを示した。これは，対数関数は,「とても遅く変化する」代表的な関数なため,「言語はとても安定してるが日々少しづつ変化する」という直観を数理的により厳密に表現したものと考えられる（ここでの使われ方の変化

は，１日でのブログや新聞での記事数や Wikipedia のページ閲覧数など単純のキーワードのカウントの時間変化である)。

　また，この現象を長期的な記憶をもつ時間に関する確率モデルでモデル化することにより， 社会的記憶との関係性を示唆した．具体的には，十分定着した単語の記憶の強さは，安定（定常）と不安定（非定常）の境界の強さにあたり，とても特別な記憶の強さということ示した。

【本研究の学術的寄与】

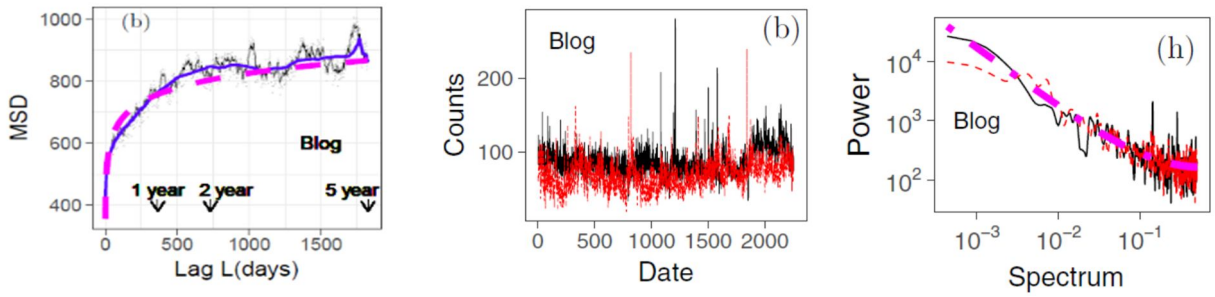
（あ）物理学： 物理学において，対数拡散は，Ultraslow diffusion と呼ばれ、数学モデルとして研究・予言されてきたものだが、現実世界での観測例は、自然現象や社会現象を問わずほとんど知られておらず，その例である。

（い）言語科学： この現象は、ブログ、新聞、英語・フランス語・中国語・日本語の Wikipedia page view など、単語・媒体・言語の種類等の系の詳細によらず観測できる．新たな言語法則かもしれない。

（う）数理科学： ブログにおける現象を説明するために用いた「べき乗的な記憶をもつランダムウォークモデル」は，非整数微分方程式に関連する。特に，1/2 階微分方程式に対応する．ちょうど 1/2 になる珍しい例である。

(え)情報学応用： 「十分定着した語」からのかい離を調べることで「新語研究」や「異常値検出」など興味ある ”特別な現象 ” の精密計測に役立つと考えられる。

図 1「立場」の件数時系列にみられる対数拡散（黒細線：データの平均二乗変位[基準日からＬ日後の平均的な件数変化]、青太線：データの移動平均，桃色破線：対数関数）．中図は対応する「立場」の件数の時系列データ（縦軸：国内ブログ単語出現数/日，横軸：日数）．黒：実データ、赤：研究モデルの数値シミレーション．右図は対応する時系列のスペクトル密度．桃色：モデルの理論値[2].



b) 新語の普及に基礎研究

　ブログにおける新語の普及の研究：日本中のブログにおける新語を体系的に収集し，出現頻度の時間変化の数量的な研究を進めた。結果，２つパラメータ相違で多くの新語時系列が体系的に説明できる簡単な微分方程式があることを明らかにした。今後，このパラメータと新語のキーワードとの定性的な性質の関係をつきとめ，論文成果化を行う予定である。また，現在研究では，普及の初期から普及後の件数を予想することは基本的に困難が多そうなことが示唆されている（例えば，初期の普及する速度が速い語は，増加期間が短く，普及速度からの普及後の予測は，困難な可能性があること，流行終盤で減速するものでなく，加速するものもあり，それが予測を困難にすることなど)。

加えて，流行状態を言語的な定量化指標の構築を目指す研究として，共起語時系列の時間的な規格化の研究を行った。結果，平均的な記事長での時間変化での件数の補正の必

要性を明らかにした。　また，関連してブログデータからみられる社会的記憶の研究については，高大連携事業において，高校生との研究もおこなった。

c) 明治大正昭和の過去新聞データに関する研究

情報システム研究機構の人文学オープンデータセンターとの歴史日本語文書の OCR 研究チームとの共同研究を開始した.それに関連して，専用ソフトにより暗号化された新聞画像データの朝刊約１００年分に関して画像解析可能な pdf データへの変換を完了した。また，OCR の前処理にあたる新聞のレイアウト解析等の基礎的研究にも進展があった。明治大正昭和の過去新聞データは，より長期的な普及現象の解析に役立つと期待されていた。

d) その他の研究

現実社会とブログのような Web データの関連性を明らかにする研究として Web 賃貸募集データと現地調査データの類似性や相違を明らかにする研究を行った．Web インテリジェンスとインタラクション研究会で途中報告を行い，スタートアップ賞と特別賞の２つの賞を獲得した。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Hayafumi Watanabe	４．巻 98
２．論文標題 Empirical observations of ultraslow diffusion driven by the fractional dynamics in languages	５．発行年 2018年
３．雑誌名 Physical review E	６．最初と最後の頁 12308
掲載論文のDOI（デジタルオブジェクト識別子） 10.1103/PhysRevE.98.012308	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 渡邊隼史	４．巻 30
２．論文標題 大規模日付つき言語データに観測される対数拡散と分数階微積分－十分定着した単語の使用はどのように安定しているか？－」	５．発行年 2020年
３．雑誌名 応用数理	６．最初と最後の頁 10-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.11540/bjsiam.30.2_10	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計18件（うち招待講演　2件／うち国際学会　6件）

１．発表者名 Hayafumi Watanabe
２．発表標題 Statistical modeling of logarithmic diffusions in word counts time series in nation-wide language data sets
３．学会等名 Conference Data Science, “Statistical modeling of logarithmic diffusions in word counts time series in nation-wide language data sets”, Statistics and Visualisation (DSSV 2019)（国際学会）
４．発表年 2019年

１．発表者名 Hayafumi Watanabe
２．発表標題 Statistical properties and modeling of stable-like word count time series in nation-wide language data
３．学会等名 Young Business and Industrial Statisticians Workshop on Recent Advances in Data Science and Business Analytics（国際学会）
４．発表年 2019年

１．発表者名
渡邊隼史

２．発表標題
Webの書き込みデータにみる数理構造と社会的な記憶―社会に十分定着した語は日々どのくらい変化しているか？―

３．学会等名
計算社会科学とその周辺セミナー（招待講演）

４．発表年
2019年

１．発表者名
渡邊 隼史，一藤 裕，鈴木 雅人，山下 智志

２．発表標題
Web不動産データを用いた空物件が入居されるまでの期間に関するデータ特性を考慮した統計モデリング

３．学会等名
2019年度人工知能学会全国大会

４．発表年
2019年

１．発表者名
渡邊 隼史，一藤 裕，鈴木 雅人，山下 智志

２．発表標題
Web不動産データを用いた空物件が入居されるまで期間の確率モデリング ―物件の特性からどこまで入居の予測が可能か？

３．学会等名
第１４回 Webインテリジェンスとインタラクション研究会

４．発表年
2019年

１．発表者名
H. Watanabe

２．発表標題
Empirical observations of ultraslow diffusion in languages: Dynamical statistical properties of word counts of already popular words

３．学会等名
Conference on Complex Systems 2018 (CCS 2018)（国際学会）

４．発表年
2018年

１．発表者名
H. Watanabe and Yu Ichifuji and Masahito Suzuki and Satoshi Yamashita

２．発表標題
Multivariate analysis of the occupations of rental rooms by using the housing information web site data

３．学会等名
International Workshop on Data Science 2018（国際学会）

４．発表年
2018年

１．発表者名
渡邊 隼史 一藤 裕 鈴木 雅人 山下 智志

２．発表標題
Web不動産データを用いた空物件が埋まる遷移に関する多変量解析

３．学会等名
2018年度人工知能学会全国大会

４．発表年
2018年

１．発表者名
渡邊隼史

２．発表標題
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―ベンチャー企業での開発―

３．学会等名
テキストマイニング 2018

４．発表年
2018年

１．発表者名
渡邊隼史

２．発表標題
日本語のブログのおける形容詞の書き込み頻度時系列の確率的特性とイベントのインパクト計量への応用

３．学会等名
日本行動計量学会 第46回大会

４．発表年
2018年

１．発表者名
渡邊隼史

２．発表標題
十分社会に定着した単語の使用は日々どれくらいづつ変化してるか？－一国規模の単語使用頻度時系列における対数拡散－

３．学会等名
経済・社会への分野横断的研究会

４．発表年
2018年

１．発表者名
渡邊隼史

２．発表標題
時間付きテキスト上に観測される超慢拡散（対数拡散）

３．学会等名
平成30年度統数研研究会「社会物理学の新展開」

４．発表年
2019年

１．発表者名
渡邊隼史

２．発表標題
様々な単語頻度時系列に共通して観測される対数的な拡散

３．学会等名
第二回計算社会科学ワークショップ

４．発表年
2018年

１．発表者名
渡邊隼史，佐野幸恵，高安秀樹，高安美佐子

２．発表標題
ブログにおけるキーワードの書き込み時系列の物理学的観点での解析

３．学会等名
計量国語学会第六十一回大会

４．発表年
2017年

１．発表者名
渡邊 隼史

２．発表標題
ブログ上の形容詞時系列アンサンブルのランダム成分の特性を利用した一国規模の社会イベントのインパクトの計量

３．学会等名
統計関連学会連合大会

４．発表年
2017年

１．発表者名
H. Watanabe, Y. Sano, H. Takayasu , M. Takayasu

２．発表標題
Statistical properties of fluctuations of time series representing appearances of words in nationwide blog data and their applications

３．学会等名
International Conference on Computational Social Science（国際学会）

４．発表年
2017年

１．発表者名
H. Watanabe, Y. Sano, H. Takayasu , M. Takayasu

２．発表標題
The probability distributions and the fluctuation scalings of the time series of key-word counts in nation-wide blog data

３．学会等名
Econophysics Colloquium 2017（国際学会）

４．発表年
2017年

１．発表者名
渡邊隼史

２．発表標題
具体例から考える”整っていない”大規模データの解析 －誤った解析を減らし，少しでも明瞭な結果を得るために－

３．学会等名
ネットワーク科学セミナー2017統計数理研究所（招待講演）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関