研　究　報　告　書

「マルチスケール社会データに対するモデリング統合技術の開発」

研究タイプ：通常型
研究期間：　平成26年 10 月～平成30年3月
研　究　者：　山田　健太

1．　研究のねらい

コンピュータの発達に伴い高度情報化社会となり身の回りの様々な情報(金融市場での取引情報、ブログやTwitterなどインターネット上での書き込み、電車の乗車記録など)が電子化されるようになった。これらの高頻度高精度情報(ビッグデータ)を詳細に解析することにより、これまでは観測が困難であったため解明が難しかった人々の集団行動に関する経験則を高い精度で確立できるようになった。また、ひとつのデータセットに限らず、個人の行動といったミクロスコピックなデータと経済指標などマクロスコピックなデータを組み合わせたマルチスケール社会データの解析やモデリングは、より深く人間の行動と社会現象の関係を解明する上で重要である。

本プロジェクトでは、SNS(ブログやTwitter)、金融市場、POS(Point Of Sales)データなど様々なビッグデータを定量的に解析することにより、これまでは観測が困難であった人間の集団行動に関する経験則を確立し、それらの発生機構を、時系列モデルや人間の行動を単純化したエージェントベースモデルによって明らかにする。モデルは仮定やパラメータが極力最少になるミニマルモデルでありながら主要な経験則を全て満たすように構築することで、モデルの持つパラメータと出力される結果の対応関係が明確に分かるようにする。これにより、ミクロスコピックな人間の行動とマクロスコピックな集団現象の関係が明らかになる。

現実のデータから観測された統計則を再現するシンプルなモデルを構築できると、このモデルを応用して様々な状況をシミュレーションすることが可能となる。例えば、安定した金融市場の構築、そして誤情報やフェイクニュース(虚偽報道)拡散の制御へ応用が考えられ、感と経験に頼らない科学的視点からの問題解決を目指す。このように蓄積されたデータを科学的に分析・モデリングし、社会活動に還元する知識循環型の社会システムの構築は、実社会からの期待も大きく21世紀の重要なテーマである。

2．　研究成果

（1）概要

(A)SNS(ブログや Twitter)、(B)金融市場、(C)POS(Point-of-sales)のビッグデータを解析することにより、これまでは観測が困難であった人間行動からも普遍的な経験則が観測されることを示した。また、時系列モデルやエージェントベースモデルを用いて、それらを再現するミニマルモデルを構築することで、観測された現象の発生機構を明らかにした。特に、マルチスケールの解析やモデリングに着目し人間行動というミクロスコピックな視点と流行や経済現象などマクロスコピックな現象を結ぶ統計物理学的アプローチを行い、複雑な相互作用が想定される人間行動や社会現象に対してもこのようなアプローチは有効であることを示した。

また、データ解析やモデリングで得られた結果や知見を応用し、流行や誤情報(フェイクニュース)の大規模伝播の予兆検知などを行った。

(A)SNS

本プロジェクトで特に力を入れたテーマであり、【単語出現頻度のパターンとその再現】、【誤情報拡散の解析】、【流行検知システムなどの開発】、【ブログデータと経済指標の相関解析】と幅広く解析とモデリングそして応用を行った。
最初にブログや Twitter などに現れる単語の出現頻度に4つの典型的パターン(流行語など)があることを発見し、エージェントベースモデルによってそれら全てを再現することに成功し、ミクロスコピックなエージェント(ブロガー)がどのように行動すると、流行というマクロスコピックな現象が発生するのかを明らかにした。

(B)金融市場

主に、各ディーラーの買い注文と売り注文の集合情報である板情報を用いて学術的側面の強い【板情報モデルの構築】、【コンピュータトレーダーと裁定機会の関係性の解析】と実践的側面が強い【取引コスト推計モデルの構築】を行った。

(C)POS

コンビニエスストアやスーパーマーケットにおける商品発注量の最適化を目的とし【POS データに基づく購買シミュレーション】を行った。特に、ある商品が欠品した際の代替商品の効果や廃棄コストを考慮したシミュレーションを行った結果、提案モデルは、毎日一定量を発注する単純な定量発注モデルよりも利益が向上することを示した。

(2)詳細

(A)SNS

ブログや Twitter などの SNS には食事や趣味などの日常的話題から政治的話題まで様々な話題が書き込まれるが、その内容は主に著者の関心事が中心と考えられる。例えば、図.1の時系列は2014年2月から4月末までのブログ中に「増税」を含む割合である。増税開始日(4/1)に向けて急激に増税に関する書き込みが増えていることが分かる。また、増税前に家電を買ったなど、駆け込み需要をあらわすブログが数多くあり、増税前の世間の雰囲気を定量的に把握することができる。

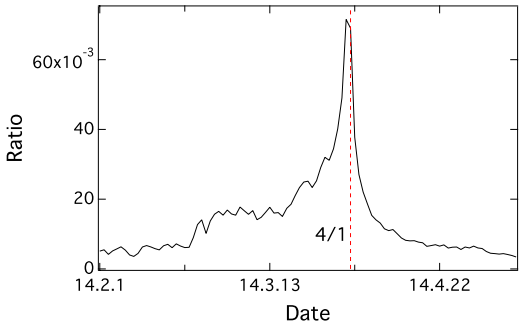

図.1 ブログ上での"増税"という単語の出現頻度

【単語出現頻度のパターンとその再現】

50万ブロガーが投稿した約1億3千万の日本語ブログの解析を行った。日次の単語の出現頻度を観測すると(i)日常語、(ii)流行語、(iii)ニュース語、(iv)季節語の4つの代表的パターンが存在することが分かり、これらのパターンは指数関数や冪関数などによって特徴付けることが可能であることを示した。また、上記4つの組み合わせにより任意の単語の出現頻度を表現することが可能である。

次に、ブロガーの行動を単純化したエージェントベースモデルによって上記のパターンを全てシミュレーションによって再現することに成功し、それぞれのパターンを再現するためにはエージェントがどのような特性を持てば良いかを検証することにより、エージェント(ブロガー)の行動というミクロスコピックな視点と流行の創発やニュースの忘却などのマクロスコピックな現象の関係を明らかにした。

【誤情報拡散の解析】

東日本大震災直後の約 370 万ユーザーが投稿した約1億8千万ツイートを用いて、震災後に発生したコスモ石油の爆発に伴い、有害物質を含む雨が降るという誤情報の拡散、及び、訂正情報の伝達による誤情報拡散の収束について解析を行った。誤情報の拡散は、近年の SNS の急速な発展に伴い SNS が存在しなかった時と比較して急速にそして大規模に広がることが懸念されるため、フェイクニュースの拡散と合わせて大きな社会問題となっているがどのように対応すれば良いかは未解明である。

データ解析の結果、コスモ石油の爆発に伴う誤情報の拡散は、(ii)流行語と(iii)ニュース語の特性を強く持ち、また、浦安市による、「有害な物質が降ることはない」という公式情報が広まると急速に誤情報の拡散が抑えられた。そこで、これらの効果を加えたエージェントベースモデルによって誤情報拡散の観測事実を高い精度で再現した(図.2)。また、このモデルを用いて、仮に浦安市の公式情報が 2 時間早く発表された場合、誤情報の拡散がどの程度抑えられたかを検証した。その結果、誤情報の拡散を抑えるためには誤情報の拡散を素早く検知し、行政機関などが訂正情報を出すことが重要であることが示唆された。

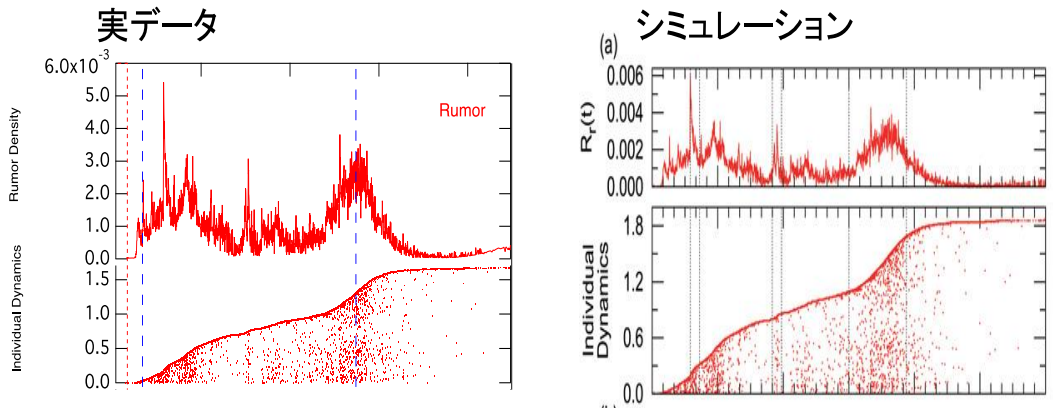

図. 2 Twitter上でのコスモ石油の爆発に伴う誤情報の拡散過程の可視化(左)とエージェントモデルによる再現(右)。 (論文発表の[1]を編集し作成)

【流行検知システムなどの開発】

任意の時系列を(i)日常語、(ii)流行語、(iii)ニュース語、(iv)季節語に分類するアルゴリズムを開発した。また、流行語の出現頻度が指数関数的上昇をするという特徴を用いて、流行の初期段階で検知するアルゴリズムを構築した。誤情報の拡散も流行語の伝播と類似した性質を持つため、このアルゴリズムは、流行の早期検知だけでなく誤情報の早期検知にも適用可能である。

【ブログデータと経済指標の相関分析】

ブログ書き込みには人々の景況感も反映されていると考えられる。例えば、下の図の時系列はブログ上での「不景気」という単語の出現頻度と内閣府より発表される景気動向指数(CI)であるが、2008 年のリーマンショック後に不景気の出現頻度が急増し、景気指数は急落しており2つの変量の間には負の相関があることが確認できる。

図.景気動向指数(上)とブログ上での”不景気”という単語の出現頻度(下)

ここでは、日経シソーラスの経済・産業[B] にある 1352 単語から、景気動向指数に対して強い連動性を持ち、かつ多重共線性を避けて独立性が保証される単語の自動抽出アルゴリズムを構築した。その結果、不景気、外国銀行、基礎的財政収支、小遣いなどの単語が自動抽出され、これらの単語の出現頻度時系列を合成した時系列によって、景気動向指数(CI)をよく再現するモデルの構築に成功した。一般的に景気動向指数など政府が発表する指数は、データの集計に時間がかかるため、発表までにタイムラグがあるが、ブログデータを用いた出現頻度の観測は毎日行うことが可能ため、速報性が高いメリットがある。また、この手法は極めて汎用性が高く、景気動向指数に限らず、他の任意の指標に対して、その指標を説明するのに適切な単語の自動抽出が可能である。

(B)金融市場

【板情報モデルの構築】

金融市場は取引履歴が膨大かつ非常に重要なことから早くから電子化が行われたため、為替レートや株価に関する経験則(価格変動の分布の冪乗則やボラティリティーの長時間相関など)はすでに確立されている。また、近年では為替レートや株価などの時系列情報だけではなく、ディーラーが金融市場に発注した買い注文や売り注文の集合である板情報も解析可能となった。

この、板情報から観測される統計的性質(指し値注文が入る位置が指数分布に従うなど)を発見した。また、この性質を有する板情報モデルを構築し、為替レートや株価から観測される経験則を再現した。これらの知見はディーラー注文やキャンセルなどの行動と金融市場の安定性と密接に関わっており市場の安定性を議論する上で非常に重要である。

【コンピュータトレーダーと裁定機会の関係性の解析】
　21世紀に入り、コンピュータプログラムによって自動売買を行うコンピュータトレーダー(AIトレーダー)が急増している。このコンピュータトレーダーの参入によって市場がどのように変化するかは、21世紀の金融市場を議論する上で必須である。そこで、コンピュータトレーダーと裁定機会の関係性の解析を行った。
　裁定機会はリスクなしで利益を得る取引であり、為替市場からはネガティブスプレッド裁定機会や三角裁定機会が観測され、取引コストや部分約定のリスクを考慮しても、裁定機会が存在することが明らかになった。また、裁定機会の発生確率とボラティリティー、取引頻度、コンピュータトレーダー数との相関関係を明らかにした。

【取引コスト推計モデルの構築】
　取引コストの主な構成要素は手数料、ビッドアスクスプレッド、マーケットインパクトである。手数料、ビッドアスクスプレッドの見積もりは比較的簡単であるが、マーケットインパクトは市場安定・不安定の状態や取引参加者の構成など様々な要因があるため見積もりが難しい。また、為替市場や株式市場において、年金の運用などで大口取引を行うと自分の買い(売り)注文によって価格を大きく動かしてしまい、より高い(安い)価格で買う(売る)ことになり、取引コストが嵩んでしまう問題点がある。また、大量の注文は市場を不安定化させる可能性がある。そこで、東証の TOPIX500 に含まれる各銘柄に対して、取引コストを算出するアルゴリズムを構築した。また、銘柄間で取引コストが優位に異なるかを検定したところ、統計的に優位に異なることが分かり、金融市場で取引を行う際に取引コスト推定が重要であることを示した。

(C)POS
【POS データに基づく購買シミュレーション】
　コンビニエスストアやスーパーマーケットにおいて商品発注量の最適化は、利益に大きな影響を与えるためもっとも重要な業務の一つである。しかし、コンビニエンスストアには約3000点もの商品があり、ここの商品の発注を店員が管理するのは非常に負担である。そのため、前日と同じ量を発注するダラダラ発注や数値として如実にあらわれる廃棄を避けるため、発注量を減らすなど(この場合、機会損失が発生する恐れがある)、損失をコントロールするという観点からは必ずしも適切な発注業務がなされてはいない。
　そこで、コンビニエンスストアの POS(Point-of-Sales)データと店舗モデル、欠品時の行動を考慮した顧客モデルを構築して、コンビニエンスストアにおける商品発注と顧客購買行動のシミュレーションを行った。基本的な解析として、店舗が定量予測発注を行う場合の粗利益最大点をシミュレーションによって算出し、また廃棄コストと欠品による購入機会損失がトレードオフの関係にあることを確認した。そして、ある商品が欠品した際に、代替品として購入される商品があることを実データから確認を行い、その効果を考慮したシミュレーションを行うと、単純な定量発注よりも利益が向上することを確認した。

3.　今後の展開

・社会実装とフィードバックによるモデルの精緻化

本プロジェクトで流行の検知システムは、既に企業で実用化されているが、社会でのより広い利用を目指して、現在利用されている現場からの感想や意見をフィードバックすることでモデルやアルゴリズムの精緻化を行う。

・テキストの内容分析を加味した誤情報・フェイクニュースの拡散検知への応用

【誤情報拡散の解析】では誤情報の拡散過程を特徴付けた。誤情報の検知をより早く、そして高い精度で行うためには、テキストの内容分析と組み合わせることが有効でると考えられるので、これまでの結果と融合させた解析を行う。

・マルチスケール×マルチフィールド社会データに対するモデルリング統合技術の開発

本プロジェクトではミクロ(人間行動など)からマクロ(経済現象など)まで様々なスケールのデータに着目したデータ分析とモデリングを行った。次のステップとしては、SNS×金融、SNS×POS(マーケティング)など複数のフィールドにまたがる解析を目指す。例えば、2013年には AP 通信の Twitter アカウントがハッキングされ、「ホワイトハウスにて爆発が2回あり、オバマ大統領が負傷した」という偽の情報によって、アメリカの株式市場は一時的に急落した。この現象を深く分析、理解するためには、SNS のデータから偽情報がどのようにソーシャルネットワーク上を拡散したか、金融市場のデータから偽情報の後、売り注文や買い注文がどのように変化したかを定量的に解析する必要がある。複数のフィールドやスケールに着目した解析の難易度は、一つのデータセットの解析と比べるとはるかに複雑で難易度も高いが、これまでのマルチスケールに着目した研究を礎とすれば可能であると期待される。

4.　評価

（1）自己評価

（研究者）

ミクロマクロを結ぶという視点から幅広くマルチスケールのビッグデータ解析とモデリングを行った。このように、データやモデルの階層構造を強く意識した俯瞰的なアプローチは、適用範囲が広く、次世代のビッグデータ解析において必須であり、本プロジェクトは、そのさきがけとなった。特に誤情報の拡散は、フェイクニュースの拡散が大きな社会問題となっている中で比較的早くから研究に取り組み、伝播過程の特徴を明らかにすることができた。また、テキストの内容ではなく、誤情報拡散パターンから検知するという試みも新規性があったと思われる。

企業の方々とディスカッションを行い、現場のニーズを受けることによって新たな研究テーマを創造し、研究結果を現場にフィードバックすることによってニーズの洗練と新たなニーズの発見につなげる「ニーズ・研究成果循環型アプローチ」を行ったが、これによって、流行の検知に関する研究成果の一部を社会実装することができた。その中で、2015年8月23日～8月26日までの4日間、米国シリコンバレーにて現地企業、大学を訪問し研究紹介およびディスカッションを行った、海外ショートビジットの経験が非常に活きた。特に研究成果の実用化を考える際には、最初は対象を絞り、その後で一般化する方がよいというアドバイスや試

作品を早く作って、早く失敗し何度もトライするのが良いというアドバイス(これらの内容はサイトビジット前にも聞いたことはあったが、スタンフォード大学のビジネススクールなど、社会の問題解決やイノベーションについて考える場所を見学することでその重要性を実体験した)が、比較的スムーズに社会実装につながった要因であると考えられる。

反省点としては、購入したデータの下処理やデータベース構築なども全て自分で行ったため、作業に非常に時間がかかった。データの下処理などに関しては研究補助者を雇用することも検討したが、なかなか適当な人材が見つからず断念したが募集の掲載範囲を広げるなどもう少し工夫が必要であった。

2016 年 8 月 24 日～26 日の日程で国際会議 Asia-Pacific Econophysics Conference 2016 -Big Data Analysis and Modeling toward Super Smart Society- (APEC-SSS 2016)を主催した。会議では、経済物理学、社会シミュレーション、情報科学、統計数理学、経済学などの第一線の研究者が一堂に会し互いの垣根を超えて議論を行い、諸分野における成果や課題を共有することにより、新たなアイディアが創発され、分野を超えた共同研究につながる研究者間のネットワーク構築を目指した。会議には当初の予定を上回る計 120 名(一般:75, 学生:45)の参加者があった。このような国際会議を主催するのは初めての経験であり、準備に多くの時間を費やしたが、参加者からは好評をいただき非常に良い経験になった。また、著名な先生を招き研究者間のネットワークが構築できたので、この新たな分野を日本が牽引するための機会として有効であった。

上記のように、基礎研究、社会実装、国際会議主催と幅広く精力的に活動し、成果をあげることが出来たので、当初の目標を達成できたと考える。

(2)研究総括評価(本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った。)

(研究総括)

高度に情報ネットワークに依存するようになった人間の集団行動に潜む法則性を膨大なデータの解析から導き出そうとする野心的な研究である。具体的に、SNS、金融市場、POSなどのデータを解析して社会科学的にも面白い結果を出している。また、数学的なモデルのシミュレーションと比較することで、経験則の妥当性を検証している。社会ビッグデータ解析による社会の動向の分析技術として、ユニークな成果として高く評価できる。今後、社会現象の原理の説明と連携できれば、Society5.0時代の社会科学を変える大きな指針を与えることにつながる。従来の研究領域の枠にとらわれず、社会科学領域の研究者や他の分野の研究者と連携して、新たな知見を生み出す活動に発展することを期待する。また、経済物理学に関する国際会議を主催し、この分野の指導的な研究者としての地位も確立している。

5. 主な研究成果リスト

(1)論文(原著論文)発表

[1] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu, "Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study", PLoS ONE **10**, e0121443, 2015

[2] 松村　直樹, 和泉　潔, 山田　健太, "POS データに基づく欠品時の顧客行動を考慮した小売店舗の購買シミュレーション", 人工知能学会論文誌 Vol.31, 2016

[3] Takuma Torii, Kiyoshi Izumi, Kenta Yamada, "Shock Transfer by Arbitrage Trading:Analysis Using Multi-Asset Artifcial Market", Evolutionary and Institutional Economics Review **12**, 395, 2016

[4]Kenta Yamada and Takayuki Mizuno, "Relationships between market impact characteristics and order book properties", Proceedings of 2017 IEEE International Conference on Big Data

[5] Kenta Yamada, "Detecting two types of seasonal word using simple autocorrelation analysis", Proceedings of 2017 IEEE International Conference on Big Data

(2)特許出願

研究期間累積件数:0 件

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

著作物:

和泉潔, 斎藤正也, 山田健太(4 章担当), マルチエージェントのためのデータ解析(コロナ社, 2017 年)

受賞:

山田健太, 高安秀樹, 高安美佐子, The 36th JSST Annual Conference International Conference on Modeling and Simulation Technology, Outstanding Presentation Award

学会発表:

1. [invited] Kenta Yamada, Hideki Takayasu and Misako Takayasu, (2017, October) "Simulations of word popularity dynamics observed from large scale social data", The 36th JSST Annual International Conference on Simulation Technology(JSST2017), Tokyo, Japan
2. Kenta Yamada, (2017, December) "Detecting two types of seasonal word using simple autocorrelation analysis", 2017 IEEE International Conference on Big Data

招待講演:

"大規模ソーシャルデータから観測される流行現象のモデル化", 電子情報通信学会複雑コミュニケーションサイエンス研究会 (CCS) (2018 年)