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

平成　２９　年　　６　月　２１　日現在

機関番号：　６２６１５
研究種目：　若手研究(B)
研究期間：　2013～2016
課題番号：　２５７５０１３０

研究課題名（和文）大規模ウェブデータを用いたうわさと流行の伝播解析とデマ拡散の制御方法の構築

研究課題名（英文）Analyses of social contagion and construction of false rumor's control methods using large scale web data

研究代表者

山田　健太（Yamada, Kenta）

国立情報学研究所・金融スマートデータ研究センター・特任助教

研究者番号：００６０９７０３

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本課題では、ブログ・ツイッターデータから観測される流行の経験則を確立し、それらを再現するエージェントベースモデルを構築することで、エージェント間にどのような相互作用が存在すると流行が創発するかを明らかにした。また、モデルを理論的に解析することにより、流行の時間発展を記述する「流行のライフサイクル方程式」の提案を行なった。さらに、これらの結果を利用して、デマ情報の早期発見や制御の方法を構築した。

研究成果の概要（英文）：We analyzed large scale social data and confirmed empirical laws of trending words. Then we constructed agent-based model which reproduces the empirical laws. Namely we clarified the relationships between microscopic agents' behaviors and macroscopic dynamics of trending words. By analyzing the agent-based model theoretically, we introduced "life cycle equation of trending words" which characterizes the time evolution of trending words' frequency. Also we constructed early detection and control methods of false rumors.

研究分野：　統計科学

キーワード：　大規模ソーシャルデータ解析　時系列解析　エージェントベースモデル　確率過程

１．研究開始当初の背景

うわさや流行の伝播過程は古くから興味が持たれ、感染症などとのアナロジーを用いた研究が 1960 年代より行われてきたが、うわさや流行の高精度な定量的観測は難しかった。しかし、近年では、インターネットの普及により、人々が自発的にブログやツイッターなどへ日々の関心事などを投稿するようになり、これらのデータは、人々の関心を強く反映していると考えられ、うわさや流行を科学的に研究するための材料が整った。

インターネットの普及により、ニュースなど様々な情報をウェブ上から獲得できるようになったが、一方、誤情報やフェイクニュースの拡散は大きな社会的問題になっている。

２．研究の目的

(1)大規模なブログやツイッターなどのソーシャルデータを解析し、うわさや流行の伝播過程に見られる経験則を確立する。

(2)人々のブログの投稿過程をモデル化したエージェントベースモデルを構築し、実データ解析から観測された経験則を再現する事で、人々の行動や相互作用という微視的な過程と流行の創発という巨視的な現象の関係を明らかにする。

３．研究の方法

(1)30 万ブロガーが書いた約 5000 万記事のブログデータや約 370 万のユーザーによる 1 億 8000 万のツイートデータを用いて、流行した単語や誤情報などの 1 日当たりの出現頻度の時系列を観測し、その特徴を明らかにする。

(2)人々のブログの投稿過程をなるべくシンプルにモデル化したエージェントベースモデルを構築する。最初は、非常に簡単なモデルから出発し、徐々に新たな効果やエージェント間の相互作用を追加していき、どのような条件があれば実データから経験則を再現できるかを明らかにする。また、シミュレーションだけでなく、エージェントモデルを理論的に解析する。

(3) 最初に誤情報の拡散と流行語の伝播の共通点と相違点を明らかにする。次に(2)のモデルを用いてデマ情報の拡散のシミュレーションを行い、その特性を明らかにする。

４．研究成果

(i)大規模なブログデータやツイッターデータを解析した結果、1 日あたりのある単語の書き込み数の推移には大きく分けて、日常語、流行語、ニュース語、季節語の４つの代表的なクラスがあることが分かった。

例えば、日常語の「いまに(いまに何々するだろうのように使う副詞)」の出現頻度の時系列は、平均値の周りで揺らぐ定常的な振る舞いをする。

次に、流行語の例としては、空気が読めないの略語として 2007 年に世間で流行した「KY」などがあり、出現頻度の時系列は、指数関数的な上昇と下降を持つという特徴がある。そして、多くの流行語の下降後の出現頻度は、完全に 0 に収束するわけではなく、日常語と同じように一定の割合で使われる。

また、震災後のコスモ石油の爆発後に、人体に有害な雨が降るという誤情報の拡散や訂正情報の拡散もこの流行語のクラスに属する。

３番目のニュース語に分類される単語には「マイケルジャクソン」がある。マイケルジャクソンが急死した直後に書き込みが急上昇し、その後べき関数で減衰する。

最後の季節語の例としては「海の日」がある。海の日の出現頻度は海のに向かってべき関数で上昇し、べき関数で減衰する。

(ii)(i)で観測された経験則を再現するエージェントベースモデルを構築した。最初に N 人のユーザーがいる SNS 空間を考える。各ユーザーは確率 p で記事を投稿し、確率 q_i で単語 i を投稿する。

p, q_i 共に一定の場合は 1 日あたりの単語 i の書き込み数は日常語と同じように、平均値の周りで揺らぐ。流行語のように指数関数的な上昇や下降を再現するためには、エージェントに基底状態(Ground state)、励起状態(Excited state)、終状態(Final state)の３状態を与える。

基底状態はユーザーがその単語をまだ知らない状態を表し、励起状態はその単語(話題)に強い興味がある状態を表し、終状態は、飽きた状態である。それぞれの状態での単語 i の書き込み確率を $q_{i,G}$, $q_{i,E}$, $q_{i,F}$ ($q_{i,E}>q_{i,F}\geqq q_{i,G}=0$)で与える。

励起状態は強い興味を持っているため書き込み確率が高く、終状態は飽きた状態なので書き込み確率は励起状態よりも低い。基底状態はまだ単語を知らない状態なので書き込み確率は 0 である。

また、基底状態から励起状態への遷移確率を $\alpha_i(\propto w_i(t))$,励起状態から終状態への遷移確率を β_i(=定数)で与える。ここで $w_i(t)$は t 日目の単語 i の書き込み数である。つまり、単語 i の書き込み数が多いほど、基底状態(単語 i を知らない)のユーザーが単語 i を知り、励起状態へ移動することを表す。このモデルは感染症の伝播をモデル化した SIR(Susceptible-Infected-Recovered) モデルと非常に近い構成であるが、SIR モデルにはブログを書き込む過程はない。また、SIR モデルでは Infected のエージェントのみ Susceptible の状態のエージェントを Infected へ遷移させる感染力を持つが、今回の提案モデルでは、$q_{i,F}>0$ の場合、終状態でもブログ中に流行語を書き込む可能性を持つので、基底状態のエージェントを励起状態へ遷移させる感染力を有する。

この、３状態の設定を加えると、「KY」のよ

うな流行語を再現できる。また、モデルのパラメータはグリッドサーチによって、実データとシミュレーションの書き込み数の距離が最小になるように、最適なパラメータを選んだ。

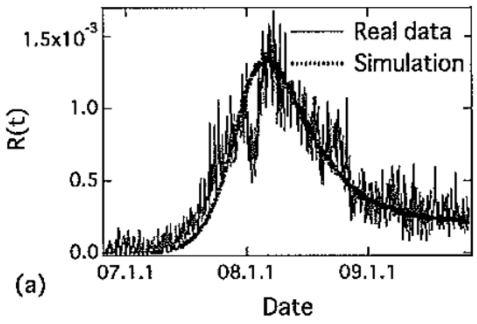

図１．エージェントシミュレーションによる「KY」の出現頻度の再現。縦軸は１日あたりのブログ全数で「KY」の出現頻度を規格化しており、「KY」を含む割合である。[１]より作成。

さらに、このエージェントモデルに対するマスター方程式から、各状態の人数に対する平均値の時間変化を連立常微分方程式で記述した。そして、この連立常微分方程式の理論解析から、流行語の流行期、衰退期、定常期という流行語のライフサイクルを記述する、流行語のライフサイクル方程式を提案した。

ニュース語や季節語の場合は、流行語を再現するモデルに、ニュース効果(α_i をある日(t)のみ 0 より大きい値を与えそれ以外の日は 0 とする)と締め切り効果(海の日であれば海の日に向けてべき関数でアルファの値が大きくなる)を加えることによって、それぞれの上昇部分を再現できることを示した。また、べき関数に従う減衰は、β_i をエージェントに対して指数分布で与えることで再現できることをシミュレーションと理論解析の両方で示した。この β_i のエージェント依存性は、飽き方に個性があることを表している。

(iii)2011 年 3 月 11 日の東北地方太平洋沖地震後のコスモ石油の LP ガスタンク爆発に起因する、有害物質を含む雨が降るという誤情報の拡散と訂正情報の拡散についてツイッターのデータを用いて解析を行った。

最初に 2011 年 3 月 11 日～17 日までに書き込まれた約 1 億 8 千万ツイートからコスモ石油を含む約 16 万ツイートを抽出し、さらに{傘、カッパ、有害物質、レインコート}のどれか一つでも含む単語を含む約 10 万ツイートを LP ガスタンク爆発の誤情報、訂正情報に関連するツイートとした。また、誤情報、訂正情報を分類するために、{訂正、デマ、誤情報、ガセ、チェーンメール、…}など訂正を表す単語を含む約 6 万ツイートを訂正ツイート、上記の訂正を表す単語を一つも含まない、約４万ツイートを誤情報ツイートと定義した。

誤情報、訂正情報の頻度の時間変化は、流行語と同じように指数関数的に増加した。一方、誤情報は、浦安市からチェーンメールのような事実は確認できないという公式発表後、訂正ツイートが広がり始めると急速に減少したため、誤情報と訂正情報の間には相互作用があると考えられる。

各ユーザーが誤情報を投稿する過程を(ii)で述べたエージェントベースモデルを拡張することで、誤情報の頻度時系列の再現を行なった。つまり、励起状態では誤情報を投稿し、それ以外の基底状態と終状態では、誤情報の投稿は行わないとした。そして、励起状態から終状態への遷移確率(β_i)は、(ii)のモデルでは一定であったが、ここでは $\beta_i(t) \propto w_{\mathrm{correction}}(t)$ と訂正ツイート数($w_{\mathrm{correction}}$)に比例すると設定した。感染症とのアナロジーでいうと、これは、訂正ツイートがワクチンのように機能し、誤情報を書き込む励起状態から終状態への遷移を促す効果である。さらに、基底状態から終状態への経路を加え、その遷移確率を $\gamma_i(t) \propto w_{\mathrm{correction}}(t)$ と設定した。つまり、この経路を通ったユーザーは誤情報を書き込まない。これは、感染症における予防接種の役割と同じである。

これらの効果を付加することによって、誤情報の頻度の時系列をエージェントベースモデルによって再現することに成功した。このように、仮定やパラメータを必要最小限に抑えたモデルで実データから観測された、現象を再現することは、エージェントの行動を解明する上で非常に重要である。

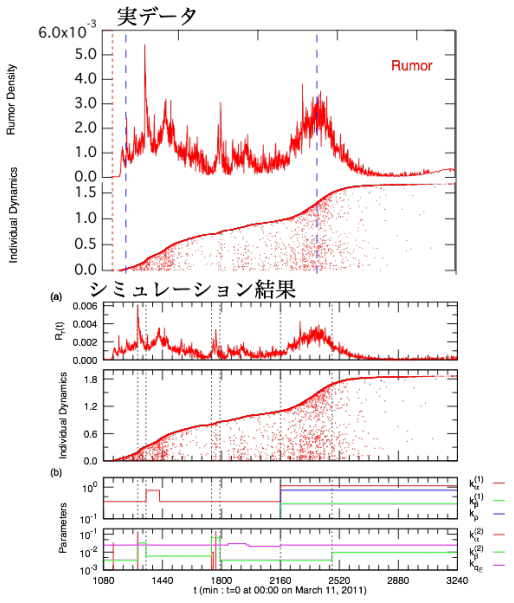

図２．誤情報の拡散に関する、実データの観測結果（上）とエージェントシミュレーションによる再現結果(下)。時系列は誤情報の頻度、下のプロットは各ユーザーの投稿過程。[２]より作成。

次に、構築したエージェントベースモデルを用いて、浦安市の公式発表が２時間前に行われた場合のシミュレーションを行なったと

ころ、デマの拡散はおよそ半分程度になることを確認した。誤情報の拡散を抑えるためには、早期に発見し、訂正を行うことが重要である。そこで、急上昇した単語や指数関数的な上昇をした単語を検知するアルゴリズムの開発を行った。

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 2 件）
[1] 山田 健太, 高安 秀樹, 高安 美佐子 サイバー空間中の口コミブームのシミュレーション(シミュレーション学会, Vol.33, No.4, pp273-278, 2014), 査読あり
[2] 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, 査読あり

〔学会発表〕（計 16 件）
① 山田健太,“ソーシャルデータに対する統計物理学的アプローチ :ブログ・twitterデータの解析とモデル化”, 早稲田大学現代政治経済研究所セミナー：政治・経済・法の計量分析, 2013.5.28, 早稲田大学(東京都・新宿区)
② Kenta Yamada, Yukie Sano, Hideki Takayasu and Misako Takayasu, “Understanding power-law growth and relaxation of collective human behaviors”, European Conference on Complex Systems(ECCS) 2013,2013.9.17, Barcelona (Spain)
③ 山田健太, 佐野幸恵, 高安秀樹, 高安美佐子, ” 人間行動におけるべき関数的成長と緩和現象に対するエージェントベースモデルの構築と解析”, 日本物理学会 2013 年度秋季大会, 2013.9.26, 徳島大学(徳島県・徳島市)
④ 山田健太, 佐野幸恵, 高安秀樹, 高安美佐子, ”ブログ書き込みに見られる社会現象の大規模観測と数理モデルの構築”, 第 5 回横幹連合コンファレンス, 香川大学(香川県・高松市)
⑤ 山田健太“ソーシャルデータの経済物理学”, 第２回数理*セミナー, 立教大学(東京都・豊島区)
⑥ 山田健太, 玉岡諒, 和泉潔, ” 大規模ブログデータを用いた書き込み数の予測手法の開発”, 2014 ソーシャルメディア研究ワークショップ, 2014.10.4, 吉池旅館(神奈川県・足柄下郡)
⑦ Kenta Yamada, “Universal Dynamics of Word Frequency Observed in Collecting Behavior on the Internet” , 2015.3.10,Singapore (Singapore)
⑧ Kenta Yamada, Ryo Tamaoka, and Kiyoshi Izumi, “Prediction of topics’ survival using large-scale social data: case of comedian popularity”, 79th Annual Meeting of the DPG and DPG Spring Meeting, 2015.3.19, Berlin(Germany)
⑨ Kenta Yamada, Yukie Sano, Kazuya Sato, Wataru Miura, Hideki Takayasu and Misako Takayasu, “Modeling transition of human interests using large scale social data”, Conference on Complex Systems 2015, 2015.9.28, Tempe(America)
⑩ 山田健太, 高安秀樹, 高安美佐子, ” 状態遷移型ブログ投稿モデルの理論解析と応用”, 2015 ソーシャルメディア研究ワークショップ, 2015.11.28, 支笏湖休暇村(北海道・千歳市)
⑪ 山田健太, 高安秀樹, 高安美佐子, ”サイバー空間中の口コミブームの解析とモデル化”, 日本物理学会第 71 回年次大会, 2016.3.20, 東北学院大学(宮城県・仙台市)
⑫ 山田健太, 玉岡諒, 和泉潔, ” 大規模ソーシャルデータを話題継続性のモデリング”, 第 30 回人工知能学会全国大会, 2016.6.8, 北九州国際会議場(福岡県・北九州市)
⑬ Kenta Yamada, Yukie Sano, Hideki Takayasu and Misako Takayasu, “Understanding lifecycle of popularity using large-scale social data”, 26th International conference on Statistical Physics, 2016.7.18,Lyon(France)
⑭ Kenta Yamada, , Hideki Takayasu and Misako Takayasu, “Blog entry model: reconstruction of popularity dynamics observed in large-scale blog data”, Econophysics Colloquium 2016, 2016.7.29, Sao Paulo(Brazil)
⑮ Kenta Yamada, Yukie Sano, Hideki Takayasu, Misako Takayasu, “Modeling transition of human interests using large scale social data”, Joint 13th Asia Pacific Physics Conference and 22nd Australian Institute of Physics Congress, 2016.12.5, Brisbane (Australia)
⑯ 山田健太, 佐藤和也, 高安秀樹, 高安美佐子, ”流行のライフサイクル方程式”, 日本物理学会第 72 回年次大会, 2017.3.17, 大阪大学(大阪府・吹田市)

６．研究組織
(1)研究代表者
山田 健太 (Yamada, Kenta)
国立情報学研究所・金融スマートデータ研究センター・特任助教
研究者番号：00609703