Twitter を利用した流行展開の分析
Analysis of growing trends using Twitter data

佐久間 雄也
Yuya Sakuma
法政大学情報科学部コンピュータ科学科
E-mail: yuya.sakuma.4n@cis.k.hosei.ac.jp

Abstract

The spread of social media has enabled people to transmit information. Accordingly, trends are arising not only from existing media but also from the Internet. Since every person can become a sender of information in the Internet, it is difficult for existing techniques to guess the transition of trends. However, if we can predict such trends in the Internet, we can use them for marketing and corporate activities. This paper presents a system that uses Internet data to analyze trends in the Internet. Especially since many people frequently transmit social media data like Twitter messages, it is possible to find people's interests from these data. By the learning of trends in Twitter data, the system finds patterns of trends, and visualizes transitions of trends. Also, this paper proposes using the concept of sociological diffusion processes. The system places high values on the early activity of a diffusion process and also on innovators who promote trends. This paper also provides the implementation of the system.

1. はじめに

近年のソーシャルメディアの台頭により，あらゆる人々が情報発信する事が可能となり，人々の関心の対象たる流行が，TV や新聞といった代表的メディアからのみならず，ネットワークからも発生するようになってきた．かつて情報発信が代表的メディアに依存していた時代は，流行とは発信する側によって「作られる物」であったが，誰しもが情報を発信する事が可能となった現代において流行とは，様々な人々が無意識的に「作る物」となっており，その発生をコントロールする事は難しく，予測なしに利用する事も難しい．一方，企業活動や社会活動において，流行を事前に察知し，対策する事は有意と言える．事前にネットワーク上に出現する流行を検知可能な手法を確立出来れば，マーケティングにおいて有用な技術足りえる．しかしソーシャルメディアの情報量は膨大で，所謂ビッグデータであり，画一的な処理だけで正確に流行を取り出し利用する事も困難である．

本研究ではソーシャルメディアの 1 つであり，日本における利用者数も多い短文投稿サイト Twitter [1]に注目する．その膨大な投稿データに対して様々な観点を用いた分析を行う事で流行を調査し，流行全てに通じるような動向を把握し，そのパターンを当てはめる形で流行の予測と抽出を行う．これにより流行発生の予兆を見通し，早期的な発見を可能とするシステムを提案する．

2. 関連研究

本研究に関連する研究として特に以下の物がある．

2.1. 普及学

社会学者 Everett M. Rogers [2]は，社会的な観点からイノベーションを考察した普及学を確立した．普及学において説明されるイノベーションの普及プロセスは，流行の拡散プロセスに近い．中でも本研究に関わり深い物はイノベーター理論である．この理論において，物事が普及する中でそれらを採用する人種は採用時期に応じて 5 種類存在するとされ，特に貪欲にイノベーションを発見し採用する初期組のイノベーターと言った人種へ普及する事で，その後の大規模な流行に繋がるとされている．

本研究ではこの理論を踏まえ，流行直前期の物事を敏感に察知し採用する人々の動向が生み出す変化を観察する事で，世間に流行し爆発的に広まっていく前に流行を検知するという目標を持っている．また，投稿情報への変化で察知するのではなく，インターネット上に個人として存在するイノベーター達を抽出可能であれば，より高度な流行の抽出も可能ではないかと考えている．

2.2. 計量時系列分析

時系列順に並ぶデータを分析し，未来の動向を予測する手法に関わってくる研究が，経済学やファイナンスの分野で用いられる計量時系列分析である．経済学者の沖本竜義 [3]は，時系列データを分析する際の大きな目的を「過去データから将来を予測する事」と定めている．この手法は，観測データの過去の動向を統計的に処理し，現在の動向から未来に取る値を確率論的に予測する．

特にこの手法は，経済発展や株価といった数値の今後を予測する為に利用される事が多いが，インターネット上における流行の動向を，1 つの言葉に対する時系列順なユーザの反応度合い，と言った数値に当てはめ直し利用できると考えられる．しかしこの手法は過去データの統計を裏付けとした予測の正確性があり，経済分析で言えば今までの経済状況や経済政策と言った確固とした情報がある．一方，人々の興味や関心で発生が揺らぐ流行においては，非常に有効な手法であるとは言い難い．その為本研究では時系列モデルの概念的な部分を利用しつつ，独自のアレンジを加えた分析手法を提案する．

Supervisor: Prof. Hiroshi Hosobe

3. 提案手法

3.1. 流行の定義

本研究において「流行」と定めるのは，生活に根付く文化的・永続的な流行ではなく，一過性であり一部の人々が熱狂するブームと称される流行である．近年の例を挙げると，子供向け商品業界で人気を誇る「妖怪ウォッチ」，大ヒット映画である「アナと雪の女王」と言った物である．特定の業界・集団の中で一過性かつ短期的な熱狂であれども，大きな影響を残した物を流行として定義する．そのような流行は多くの場合，古くからの物ではなく新たに登場したサービス，商品，物と言った「名詞」であると考え，分析の対象として判断される物を，そのような固有名詞や辞書的には未知の言葉とする．

3.2. 流行の学習

本研究の目標は，Twitter 上における様々な物に関する流行の共通点や動向を調査し，最終的には全投稿データから，関心の対象や流行物へと昇華する可能性の高い物を取り出す事にある．その準備として，流行という曖昧な対象を一定の形として理解する為，実際に流行した物を様々な観点から分析し流行をパターン化する．そこでデータ収集期間に実際に流行した物を学習用データとして利用する．しかし判断基準無しに学習用データを選定する事は難しい為，選定は主観的な手法で行う．基準としては，様々なメディアで目にする事のあった物を，ジャンル問わず用意する事とし，偏らないよう選定する．

3.2.1. 分析の基本手法

本研究における分析手法は，主に信号処理の分野で用いられる，現時点から過去一定期間のデータから推測される未来の値を予測する手法を採用する．日毎の時間軸をtと定めると，現在日時をTとした場合，Tから過去に向って一定時間を遡る$T-n$までの$n+1$日間のデータを利用し分析を行う．日毎に保持されている様々なデータを元に，Tから$T-n$までの平均値，時間軸を移動する際の変化量などの短期的な過去データを参照する．またこの期間的な分析によって，流行の動向には一定のパターンがあると考え，他の流行が見せたような短期的な動向を現時点で倣っている場合，これもまた流行の動向を見せるだろう，という推測に基づく分析を行う．

3.2.2. 分析における情報整理

分析における基本的なパラメータは言葉の頻出数である．流行語の出現量を見る事で，人々が多く投稿する程に関心があると仮定する．この頻出数を時系列順に扱う上で，今後の動向予測やデータのぶれを考慮し整理する為に最小二乗法の概念を利用する．実験データの場合，$(T-n, y_{T-n})$から(T, y_T)となる計測値が存在する時，それらを座標空間に描いた場合において，最も近い線分を導き出し，その線分の傾きを確認する事で，データの傾向を把握する事が可能となる．この計算手法については，$n+1$個のデータ群が取る最も近い直線を，

$$y = at + b$$

とした場合，傾き値であるaは，

$$a = \frac{(n+1)\sum_{t=T-n}^{T} t y_t - \sum_{t=T-n}^{T} t \sum_{t=T-n}^{T} y_t}{(n+1)\sum_{t=T-n}^{T} t^2 - (\sum_{t=T-n}^{T} t)^2}$$

として求める事が出来る．この式は 1 次式を算出する物であるが，これは 2 次式版を採用した場合，直近に人物の訃報などがあると起きやすい短期的な爆発上昇が発生すると流行とは言えないだろう言葉まで抽出される危険性がある為，より簡素な 1 次式版の採用を決定した．

3.2.3. 詳細なパラメータ分析

パラメータ分析の 1 つの指標となるのが投稿の付加情報ある．他者に対して情報を拡散する機能であるリツイートの割合や，流行語を投稿したユーザの情報をパラメータとして扱う．Twitter には様々なアカウントがあるが，著名人が利用する物や交流が広い物と，友人知人との交流にのみ利用される小規模な物とでは，情報拡散力の差は明確である．また機械的に自動投稿する BOT と呼ばれる物も数多く存在し，分析結果にぶれが生じてしまう可能性もある．そこでユーザ分類を，アカウント毎のフォロワー（追っかけ）数や，逆の意味であるフォロー（読者）数，投稿数と言った情報から，表 1 の様な 5 分類にカテゴライズし，その比率をパラメータとして利用する．

表 1　アカウント分類

著名人	フォロワー1 万人以上 and フォロー10 人以上
一般人	フォロワー100 人以上 and フォロー10 人以上
宣伝用	フォロワー100 人以上 and フォロー10 人未満
BOT	自動投稿と思しき記述体系やアカウント説明
小規模	フォロワー100 人未満 or ツイート数 300 未満

またユーザ分類だけでなく，ユニークなユーザの割合もパラメータとして加える事で，1 人が多重に投稿する事で頻出する物ではなく，様々な人に拡散され投稿されている情報が流行の兆しになると考え，利用する．

以上のようなパラメータを流行語に対する詳細な分析観点として取り入れる．学習用データの分析においては，多数用意したデータから各種パラメータが取る平均値を流行が取る平均的なパターンとして利用する．流行を抽出する実験においては，各種パラメータが平均値に近い程流行に近い存在であると判断される．

3.3. データ分析と流行の抽出

学習用データを用いた流行のパターン化を行った後は，取得した全ての投稿データから今後の流行を抽出する事が可能なシステムを構築する必要がある．主な動きとしては，投稿データから形態素解析で抽出された名詞群を元に，特に時系列順において上昇傾向が見られる物を候補語として抽出し，更にそれらの候補語について，各種パラメータとパターン化された流行の平均値を見比べる事で，より流行として確実性の高い言葉を選別する．

流行パターンとの比較については，各種パラメータが平均値から遠い程スコアを下げていく手法を取り，学習用データが見せた最大値，最小値範囲内であれば減点，範囲外であれば大幅な減点を行い，最終的に全項目がおおよそ範囲内に収まった物は流行として抽出され，その中でもスコアが高い物はより確実性が高い流行とする．

3.4. イノベーターの発見

流行抽出の他，流行という物の共通点の 1 つとして普及学で提唱されるイノベーターが本実験で発見できるかどうかも確認する．こちらは学習用データを用いるのではなく，本実験である流行の抽出を行った後，抽出された流行語を，システムが流行として判断する直前の流行初期段階で投稿しているユーザを抽出し，特に数多くの流行語で発見されたユーザを仮イノベーターとして選別し，彼らがどのようなアカウントなのかを判別する．

4. 実装

4.1. 基礎情報，及び元データの取得

本研究では，システムの実装にプログラミング言語 Java を利用した．Twitter からのデータ取得に関しては Java 用 Twitter API ライブラリである Twitter4J [4]を利用し，2014 年 8 月 1 日から 2014 年 10 月 31 日までの 3 ヶ月分の投稿データを実験用に用意した．

4.2. 形態素解析

流行となる言葉は，主にサービス名や商品名，場所名といった「名詞」である事が大半である．しかし投稿データは様々な書式・文体で記述される為，全て画一した規格として扱えるよう形態素解析を行い，名詞のみに分解する．形態素解析には，Java 用形態素解析器 lucene-gosen [5]を利用し，その辞書には IPAdic [6]を用いる．

IPAdic に登録されている名詞は，国語辞典に掲載されているような一般的な名詞しか存在しない．その為流行の対象になるであろう，様々な新語・未知語に対応出来ない．その対策として，様々な新語に素早く対応し，豊富な知識が統合された辞書として，40 万語以上の単語が登録されている Web 辞書はてなキーワード [7]のキーワードリストを形態素解析辞書に追加し，新語や未知語に対応する．はてなキーワードの単語は IPAdic 用フォーマットに変換し，解析器に組み込む形で利用する．解析時に重要となる単語コストに関しては，はてなキーワード上の単語は多くの場合新語，未知語に近い名詞であると考え，IPAdic 上の名詞よりも採用されやすくなる形でコストを設定し，その上で単語が細かく分解されないよう，文字数が長い単語程採用されやすいように調整した．

4.3. 流行データの可視化

流行語などに付随する様々な要素をデータ化すると共に，人の目で見ても分かりやすい形とする為，簡易的な可視化システムを導入した．可視化には縦棒グラフを利用し，日毎の頻出数を長さで表現し，頻出数グラフ内に，リツイート割合やユニークユーザ割合を描画する．

5. 実験

5.1. 流行のパターン化実験

データ収集期間中，実際に流行した学習用データについての傾向を分析する．図 1 は学習用データの 1 つである，サブカルチャー分野で話題となった作品の流行推移を可視化した物である．8 月中は低い数値を維持していたが，9 月中旬にかけて上昇傾向を見せ，10 月期には高い数値で推移している． 10 月付近の盛り上がり部分を流行のメイン時期と考えると，9 月中旬期に見せた徐々な上り坂傾向が流行の兆しと見て良いだろう．

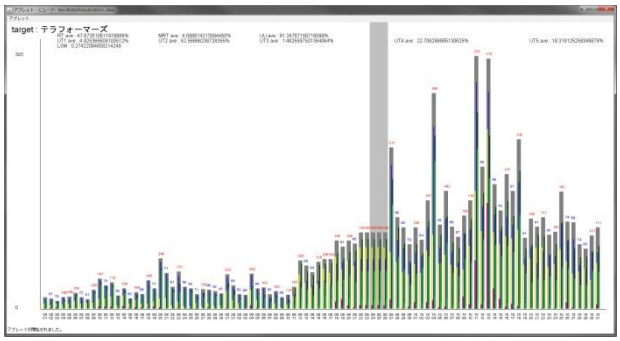

図 1　学習用データ「テラフォーマーズ」の可視化

このような傾向分析と全体的なリツイート傾向，ユーザ情報の数値などと照らしあわせて分析した結果，今回用意した流行の学習用データからは，流行の予兆期における表 2 の様な動向の傾向が見て取れた．

表 2　学習用データのパラメータ

	平均	最大	最小
リツイート	45%	73%	22%
多数リツイート	14%	26%	3%
ユニークユーザ	78%	95%	64%
著名アカウント	2%	11%	1%
一般アカウント	57%	75%	33%
宣伝用アカウント	1%	6%	0%
BOT アカウント	10%	44%	1%
小規模アカウント	19%	31%	11%
前日比上昇回数(10 日間)	5 回	7 回	3 回
前日比増減率(10 日間)	575%	3876%	104%
データの傾き	0.029	0.168	0.003

分析は流行直前期の 10 日間のデータを元に行ったが，頻出数の傾きは若干のプラスを見せ，平均して 10 日中 5 日の前日比上昇，期間中は 6 倍近い上昇率で推移する事が判明した．またリツイートが占める割合も半数近い事から，多くの流行でリツイートの効力が証明されたと言って良いだろう．ユーザ傾向も，他のユーザと繋がりが深い一般アカウントが占める割合が高く，また意外な点として BOT が占める割合も比較的多かった．これは自動投稿による物であっても，人の目に映る事で流行のきっかけとして作用する可能性も考えられるのではないかと思われる．逆に宣伝用アカウントは殆ど誤差と言える少なさで，無視しても問題ない影響であると考えられる．

以上の結果から，リツイートという機能の重要性，Twitter で交流しあう比較的活発な一般人の比率，前日比における上昇傾向の 3 点に注目し，実際のパラメータとする事が流行の兆し発見に繋がると考えられる．

5.2. 流行の抽出実験

5.1 項で算出されたパラメータと範囲を利用し，3.3 項の手法を用い 8 月から 10 月までの全投稿データから流行と思しき言葉を抽出する．単語それぞれの頻出数や最小二乗法を用いた分析による流行候補語の抽出を行い，候補語についてパラメータを利用した詳細分析を行った．

図 2 は，抽出された流行の 1 つを可視化した例である．「デスティニー」とは 9 月 9 日に発売のゲームソフトである．図では 9 月中旬付近から爆発的な広まりを見せ，その後は落ち着くもののある程度の話題として継続している事が判断できるため，流行であると見て良いだろう．

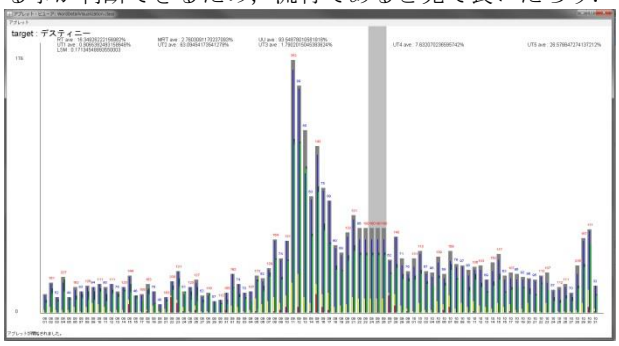

図 2　抽出流行「デスティニー」の可視化

流行の導入部分のみをパターン化した分析手法ではあるが，その後の推移を見てもある程度の話題として残る流行を抽出出来たと言える．しかし上記の流行としての話題性を持つ物の他，約 3 割程度に「大雨警報」「お彼岸」と言った，時期的に話題にはなる言葉ながらも普遍的で流行とは言えない言葉も検出されてしまった．このように精度面に関してはやや不十分な結果となった．

5.3. イノベーターの発見実験

予備実験として 3.4 項で説明したイノベーターの発見実験を行った．結果として複数の流行で見られたアカウントの詳細を確認した所，多くのアカウントは様々なニュースサイトの情報を引用し投稿するだけの物や，同様な内容をリツイートするだけの物であり，所謂トレンドリーダーたるアカウントは発見されなかった．

この結果について 1 つ考えられる点は，流行をジャンル分けせずに扱った事である．ユーザは基本的に自分の嗜好のジャンル内で交流し，流行もその中で発生する事が多く，それを跨った全体的な流行は稀有である．その為流行を一括りにイノベーターを発見しようとした場合，全ジャンルを機械的に情報収集するだけのアカウントが該当してしまい，今回のような結果になったと言える．

6. 議論

実験の結果，流行をある程度の精度で抽出出来たと言える．しかし以下のような問題点が浮き彫りになった．

6.1. 構想の問題点

問題点の 1 つに，流行の定義付けが不足であった点がある．流行と言っても，一般的に浸透した大規模な物から，特定ジャンルを知る人のみ浸透した物など，無数の種類が存在し定義も曖昧である．本研究では期間は数週間から数ヶ月程の持続と断定し，ジャンルを考慮しない学習用データの用意，平均値を使用したパラメータ設定などを行い，曖昧な流行という概念を定義付けしたが，その結果として抽出される言葉も，厳密には流行と言えない曖昧な候補語が混ざってしまった事は否定出来ない．

6.2. システムの問題点

システムの面では，1 つに形態素解析の問題があった．本研究では流行の対象を名詞・新語と仮定し，形態素解析辞書や Web 辞書を利用したが，双方ともそのまま利用した為，流行とは言えない言葉までもが対象として存在してしまった．この点に関しては，関心の対象としての言葉であるかを考慮した上での辞書再構築が必要であるが，どのような言葉が出てくるか把握できない流行に適宜対応した辞書の作成には困難が伴うだろう．また辞書の改善だけではなく，投稿文を分析し「何を対象とした文か」という観点から絞り込んでいく必要もあるだろう．

もう 1 つの問題は不十分なパターン化である．過去情報の蓄積が足りず，小規模なデータを用いた所から予測を行った点や，最小二乗法に 1 次式版だけを利用した点などが不安点である．精度向上には，より長く数年・数十年単位で蓄積された情報から統計的に分析した傾向を取り入れる事で，改善が可能ではないかと考えられる．

7. おわりに

本研究ではソーシャルメディアにおける流行の予測の為，過去に流行したと判断できる実例を学習用データとして利用する手法を提案した．高い精度と信頼性を持つシステムを構築するという点ではまだ問題点が多い事が判明した．今後の課題としては，流行という物をより深く理解・研究し，その形を掴むと言う社会学的な問題の解決や，時系列データをより様々な観点から分析し，パターン化すると言う統計学的な問題の解決が必要だろう．

文　献

[1] Twitter, https://twitter.com/.

[2] Diffusion of innovations, http://en.wikipedia.org/wiki/Diffusion_of_innovations.

[3] 沖本竜義, 経済・ファイナンスデータの計量時系列分析, 朝倉書店, 2010 年 2 月 1 日.

[4] Twitter4j, http://twitter4j.org/ja/index.html.

[5] lucene-gosen, https://code.google.com/p/lucene-gosen/.

[6] IPAdic, http://sourceforge.jp/projects/ipadic/.

[7] はてなキーワード, http://d.hatena.ne.jp/keyword/.