「マルチメディア，分散，協調とモバイル
(DICOMO2018) シンポジウム」 平成30年7月

大量ツイートの収集・分析を個人で手軽に実現可能にする方法の提案

松浦 智之[1]　當仲 寛哲[1]　大野 浩之[2]

概要：短文投稿 SNS“Twitter” は今や多くの人々に認知され，NHK を始めとした日々のニュース番組等においても，もはや Twitter やツイート（Twitter に投稿される文章）が何かという説明が省略されながら，世論を反映した情報源として引用あるいは分析されている．しかしながら，社会現象のような膨大な量のツイートを発生させる話題を分析しようとなると，既によく知られている方法では費用的にも技術的にも個人には敷居が高い．本稿では，一定の制約はあるなかでも，個人による大量ツイートデータの収集・分析を実現し得る手法を提案し，実際に，日本国内で社会現象を起こして大量のツイートを発生させた 2 つの話題に関するツイートの収集・分析を行うことで，提案手法の実用性を示している．

The Method of Collecting and Analyzing a Large Quantity of Tweets Easily for Personal Use

Tomoyuki Matsuura[1]　Nobuaki Tonaka[1]　Hiroyuki Ohno[2]

1. はじめに

短文投稿 SNS“Twitter” は今や多くの人々に認知され，NHK を始めとした日々のニュース番組等においても，もはや Twitter やツイート（Twitter に投稿される文章）が何かという説明が省略されながら，世論を反映した情報源として引用あるいは分析されている．

ツイートの分析は，そのような大きな組織のみならず，個人やそれに類する小規模な研究チームやにとっても有用である．しかしながら，世界中で日々投稿されるツイートデータの量は膨大で，それらを全て，あるいはリアルタイムに取得するために提供されている API の利用料は非常に高額である．また，データが膨大であるために，それらを取り扱う機材もしくはサービスもまた高額であり，個人にとっては敷居が高い．

本稿は，このような状況にありながらも，個人によるツイートデータの収集・分析を全く諦めるのではなく，まず制約はありながらもそれらを実現する手法を考案し，次にその手法の有効性を実証するために，日本国内で社会現象を起こし，大量のツイートを発生させた 2 つの話題に関するツイートの収集・分析を試みたので，その結果の概要を報告し，本研究が提案する手法がいわゆるビッグデータと呼ばれる研究の重要な一翼を担うことを示す．

2. 背景

2010 年台の今，Twitter は，人々のほとんど生の（誰にも内容を編集されていない）の所感や告知を，速報性のある状態で知ることができ，かつ世界規模で普及している数少ない媒体であると言える．このような特徴もあり，報道媒体等の公共性のある組織も Twitter を世論分析の情報源として活用しているが，Twitter の情報源としての有用性は，このような大規模な組織に限ったものではない．例えば自社商品のマーケティングに活用したいと考える中小企業や，世論の興味深い傾向を明らかにしたいと思っている個人やそれに類する小規模な研究チームやにとっても有用である．

そこで筆者らはまず，ツイートデータ収集・分析に関する先行事例を調査した．その中でも東京大学の鳥海の 2015 年の報告書 [1] は，ツイートデータ収集・分析に関する現状やデータ規模，収集・分析ツールの紹介など，これから

1　(有) ユニバーサル・シェル・プログラミング研究所
2　金沢大学総合メディア基盤センター / 慶應義塾大学 SFC 研究所

Twitter データ分析を始めようとしている者が把握すべき基礎知識の比較的新しい状況がまとめられていた．

2.1 無料 API による収集の制約

そして報告書には収集・分析ツールとして，TTC/TTM[2]，および TTC の機能を拡張した WTC が紹介さえていた．特に TTC/TTM は，制作者自身が紹介しているのみでも 100 を超える研究に利用されており，定番と言える．

しかしながらこれらのツールは，Twitter 社が無料で公開している “Standard search API[3]” を利用しており，第一に，過去 7 日分までしか検索でいないという制約がある．他にも，その API に課せられたアクセス頻度制限（レートリミット）への対応方法の問題により，一日あたり約 18000 ツイートを超える規模の話題を集めることが難しい．先程の報告書においても，古いツイートや大量のツイートに関しては有料で収集することが推奨されていた．

2.2 有料 API の費用問題

そこで，ツイートデータを有料で収集する方法を調査したところ，小規模な組織や個人にとっては非常に高額な費用を負担しなければならないことがわかった．

Twitter 社と戦略的ソリューションパートナー契約を結び，日本国内におけるツイートデータ販売の代理店業務を行っている NTT データが提供する API[4] の場合，過去約 1 年分のツイートデータを任意のキーワードで検索して配信する「ヒストリカルサーチ API」という比較的安価なサービスであっても，月額 45 万円以上，最低契約期間 6 か月，初期費用別途とされており約 300 万円以上の費用負担が見込まれる．

また，本研究開始後ではあるが，2017 年 11 月に Twitter 社は，このような高額な費用負担を緩和するために “Premium APIs” というサービスを発表した．一例を挙げると，Twtitter サービス開始当時まで遡れ，毎月 125 万ツイートまで取得できるプランでは月額 1899 米ドル，遡れる期間が過去 30 日という制限を付ける場合は 699 米ドルである [5]．毎月取得可能なツイート数に応じて料金は 149 米ドルから用意されており，中・小規模のツイートを収集する場合の敷居は低くなった．しかし，社会現象と呼ばれるほど大規模な話題を分析するために大量のツイートを集めたいというニーズに応えるものにはなっていない．

2.3 本研究の目的

無料で使える Standard search API には主に 2 つの欠点がある．

- 過去 7 日前までのツイートしか収集できない
- レートリミットが厳しいために大量ツイートの収集には適していない

本研究では，Standard search API を利用しつつ（すなわち個人には現実的ではない費用を発生させず），後者の欠点を克服することを目標とした．そのために，実際に社会現象と呼ばれる話題で生じるツイートデータ規模の確認，ツイートデータを収集・分析する手法の提案を行ったうえで，社会現象とされた実際のテーマに基づくツイートを収集・分析し，提案手法の有効性を検証した．

3. 「社会現象級」ツイート収集の実現可能性

3.1 「社会現象級」ツイートの規模

先に記したとおり，NTT データは日本におけるツイートデータの販売業務を行っているため，現在までのツイートデータも自由に参照できる立場にあるものと思われる．実際に同社は「イマツイ」というブログサイトを運営し，全ツイートデータを参照して得られた興味深い分析結果を公表している．

イマツイにて 2016 年 12 月 28 日に公開された記事 [6] によれば，2016 年の元日から同年 12 月 12 日までに発生した全ツイートを調べ，音楽，映画，流行語，テレビドラマという各ジャンルで話題になった語が含まれるツイート数を数えたところ，1 位が「ポケモン GO」であり，これを含むツイート数が 1929 万 8506（リツイートは除く），同様に 2 位が「PPAP」であり，1091 万 2400 であった．どちらも同年に世界的に流行し，社会現象と呼ばれるほど話題になったコンテンツであったが，それら社会現象級のツイート数は年間一千万のオーダであることがわかる．

ただし年間を通してツイート頻度が一定であったとは考えづらく，実際，同記事によれば 1 位の「ポケモン GO」は同年 7 月が最も多く，その数は 912 万 6484 と報告されている．これを 1 日あたりとして平均すればおよそ 30 万ツイートになるが，もちろん日によって差はあるはずである．仮に最も多い日と少ない日の差が 10 倍に達していたと仮定すると，この話題で 1 日で 300 万ツイート（リツイート除く）が発生した日もあったことになる．

3.2 無料 API の収集速度

一方，無料で使える Standard search API の収集速度を検証する．

Twitter 社が公開している仕様 [3] によれば，この API の 1 回呼び出しで最大 100 ツイート収集できる．そして，レートリミットによって，15 分あたり 450 回まで呼び出せる（アプリケーション認証の場合）．1 秒あたりに換算すれば，毎秒 50 ツイートまで取得できる．これを一日当たり，一か月あたり，一年あたりに換算すれば，それぞれ 432 万ツイート，約 1.3 億ツイート，約 15.8 億ツイートということになる．よって単純計算では，前述の社会現象級のツイートの発生量と比較すると，一つの話題に絞りさえすれば十分収集可能といえる．

もちろん，先程の「ポケモン GO」ツイート数の見積もりで仮定した多い日と少ない日のツイート数の差が実際には 10 倍以上である可能性や，リツイートを含めるとさらに数が増える可能性，さらに仕様書どおりの性能が長期間は持続しないなどの可能性を考慮すると，実現できない可能性もある．しかし，このように厳しく見積もれば 1 日あたりの収集限界ツイート数を超える可能性があるものの，厳しい条件が何日も持続するとは考えにくく，実際は収集できるのではないかと予想した．

4. 提案する大量ツイート収集手法

4.1 収集時点でのツイートデータ絞り込み

先程の社会現象級ツイートに関する考察により，一つの話題に絞れば無料の API で集めきれる可能性があることを確認した．そこで，全世界で発生するツイートをとりあえず全て集めるという方針ではなく，Standard search API を利用してツイートの取得段階で必要なツイートに絞り込むことを基本方針とする．これは，2.1 項で引用した報告書に記されていたものと同じである．

ただし異なる点は，ユーザ認証ではなく，アプリケーション認証を用いる点である．ユーザ認証によって Standard search API を利用した場合，そのユーザがフォローしている鍵付きアカウントの限定公開ツイートも収集でき，逆にブロックされているアカウントのツイートは収集できない等，ユーザの閲覧権限に応じて収集対象となるツイートの範囲が変化するが，大量収集の場合においてはそのような限定公開ツイートは不要であるうえに，レートリミットが厳しく設定されている（毎秒に換算すると最大 20 ツイート）．アプリケーション認証にすると，ユーザ認証ではないため，未ログイン状態で閲覧可能な範囲のツイートが取得され，かつレートリミットも毎秒 50 ツイート相当に緩和される．その結果，2.1 項のとおりの収集速度が引き出される．

4.2 ツイート ID 管理を厳密にした降順収集

ツイートの収集は，図 1 のようにして行う．

Standard search API では，指定された日付（時刻は UTC+0）またはツイート ID を起点としてツイート ID 番号の降順（過去に遡る方向）にツイートを収集できる．そこで，$n+1$ 回目の呼び出しでは，n 回目の呼び出しで収集されたツイートの中から最も小さい ID 番号から 1 を引いたものを起点にして呼び出せば連続的に集められる．ただし，レートリミットを超えないようにするため，呼び出し間隔が 2 秒になるように，呼び出し時間を調整しながら呼び出す．

これを繰り返せば，理論上最大 7 日前までのツイートが集められる．ただし，収集作業をしている間に 7 日前より古くなってしまうツイートもあり得るため，初回は 7 日前

図 1 Standard search API によるツイート収集の基本方針

の日付を起点として遡れるだけツイートを収集し，それ以上遡れなくなったところで次は 6 日前の日付を起点にして，既に取得済のツイート ID 番号に達するまで行う．長期間に渡って収集を行う場合には，収集を行った日の分まで集め終わった後，翌日の UTC+0 を過ぎるのを待ってから翌日の日付を起点にして同様の収集を行えばよい．

ツイートを過去に遡りながら収集するという基本方針も，2.1 項で引用した報告書で述べられていたツールが採用していたものと同じである．しかしながら，一定量以上のツイートを収集するにはユーザがその都度収集実行のための収集ボタンを押さなければならないうえに，レートリミットを超えないようにボタンを押すタイミングをユーザに任せていること，そして収集ボタンを使って集めた後にツイート ID が保存されないため，続きのツイートから集められないといった問題があり，大量収集に適した作りになっていなかった．本研究では，4.4 項で述べるようにツイート ID を保存して，それらの問題を解消したツイート収集プログラムを作成した．

4.3 持続性の高い収集プログラムおよびツイートデータフォーマットの実現

Standard search API では過去 7 日前のツイートまでしか遡れないため，例えば 1 年分のツイートを収集するには，ほぼ 1 年間収集作業を行わなければならない．長期間に渡って同じプログラムを使う場合に注意しなければならない問題の一つに，プログラムやデータの互換性や持続性がある．

もし収集作業期間中のある日に，例えば収集プログラムを動かしている OS をアップデートしたことが原因と疑われる理由で動かなくなったら，収集作業ができなくなってしまう．プログラムが使えなくなってしまうような環境変化は，他にも様々な要因によって起こり得る．したがって，長期間に渡る収集を行う際には特にプログラムあるいは実行環境の持続性・互換性に注意を払うべきである．

また，持続性や互換性という性質は，プログラムのみならず収集されたデータにとっても重要である．私達が収集した貴重なデータを，数十年後に後生の研究者が，当時の世論を研究する貴重な情報源として利用する可能性もある．その時，データフォーマットがすでに廃れたものになっていたら，彼らはデータを開くのに苦労したり，最悪諦めねばならないかもしれない．

以上の懸念から，収集プログラムを作るにあたり，著者らは「POSIX 中心主義」を採り入れた．

4.3.1 POSIX 中心主義

POSIX 中心主義 [7] とは，互換性や持続性に最大限の配慮をし，プログラムやデータが環境変化に強くなるようにするためのプログラミング指針である．

基本的には POSIX 文書に記されている範囲でプログラミングすることにより，多くの OS で動作する互換性，そして POSIX の改訂が非常に少ないゆえの持続性を獲得する．また，POSIX の範囲で実装できない部分に関しては，POSIX 範囲外のソフトウェアへの依存も認めるが，同一機能を持った複数のソフトウェアに対応させることを必須とし，RAID や電源二重化のような冗長性持たせ，結果として高い持続性を確保するというものである．

4.4 持続性の高い収集プログラムおよびツイートデータフォーマットの実現

以上の方針に基づき，収集プログラム「小鳥男」[8] を作成した．POSIX 中心主義に基づいているため，プログラムの大部分は POSIX 版 sh 互換のシェルスクリプトである．このプログラムは目論見どおりに多くの OS 上で動作することが確認された．また，シェルからコマンドとして実行できるため，日々の収集作業の自動化も容易に行えた．

そして，収集されたツイートデータの格納型式についても，環境変化に対する耐性が高くなるように配慮した．ツイートの投稿年月日及び時分秒に応じてディレクトリとファイルを分け，日時に基づく検索性を高めると共に，POSIX 規格で用意されている基本的な UNIX コマンドで簡単に取り扱えるよう，半角空白区切りのプレーンテキスト形式を基本とした [10].

5. 社会現象級ツイート収集・分析の検証

前節の提案手法に基づき，社会現象と呼ばれた二つの話題に関するツイートの収集・分析を実際に行った．

5.1 「バルス」ツイート

その一つは，アニメ映画「天空の城ラピュタ」[11] がテレビで放送される度に大量投稿される「バルス」という語を含むツイートの収集による検証である．

「バルス」とは，劇中で登場する呪文であり，主人公らがこの呪文を唱えるシーンが放送される瞬間に合わせ，多くの視聴者がこの呪文を含むツイートを一斉に投稿することが恒例になっている．過去にその影響で Twitter のサービスがダウンしたこともあり，ダウンを狙って投稿する者も増え，2013 年 8 月 2 日の再放送時には，その瞬間に秒間ツイート投稿数（バルス以外も含む）が 14 万 3199 を記録した [9].

このように，バルスは瞬間的に大量のツイートを発生させるイベントであり，これが収集できるかどうかも本提案手法の実用性を検証するうえで重要な指標の一つになる．

5.1.1 ツイートデータ収集

まず，対象としたバルスイベントは，2017 年 9 月 29 日放送分の天空の城ラピュタである．

そして，この検証で使用したコンピュータの主なスペックは次のとおりである．

CPU:AMD Ryzen 1700，メモリ:16GB，HDD:1TB，OS:Windows Subsystem for Linux (Windows 10)，インターネット回線:100Mbps

これは，個人向けに販売されている一般的な PC およびインターネットプロバイダで構成された環境である．

また，ツイート収集の際に使用した検索クエリは次のとおりである．

```
バルス OR "バルス祭り" OR #バルス OR
#バルス祭り OR ヴァルス OR #ヴァルス OR
barusu OR #barusu OR bals
```

このようないくつかのパターンで表記揺れが起こるものと想定し，バルスを意図していると思われるものは一通り収集されるようにした．

5.1.2 分析

ここでの分析は，収集されたツイートの毎分の数を求めることとした．NTT データ「イマツイ」が天空の城ラピュタ再放送の都度，同様のデータを公開しており，両者の収量を比較できるためである．

毎分のツイート数を数えるにあたっては，格納されているツイートデータが 4.4 項で述べたように，UNIX コマンドで扱い易い構造になっているため，図 2 のような簡単なシェルスクリプトで可能である．

```
cd /PATH/TO/TWEET_DATA/DIR
find . -name '*.txt'          |
xargs cat                     |
awk '{print substr($1,1,16)}' |
sort                          |
uniq -c > count_TPM.txt
```

図 2 1 分毎のツイート数を数えるシェルスクリプト

ちょうど年月日時分までディレクトリ分けされており，その中にあるテキストファイルが 1 ツイート 1 行になっているため，各ディレクトリの行数を数えている．

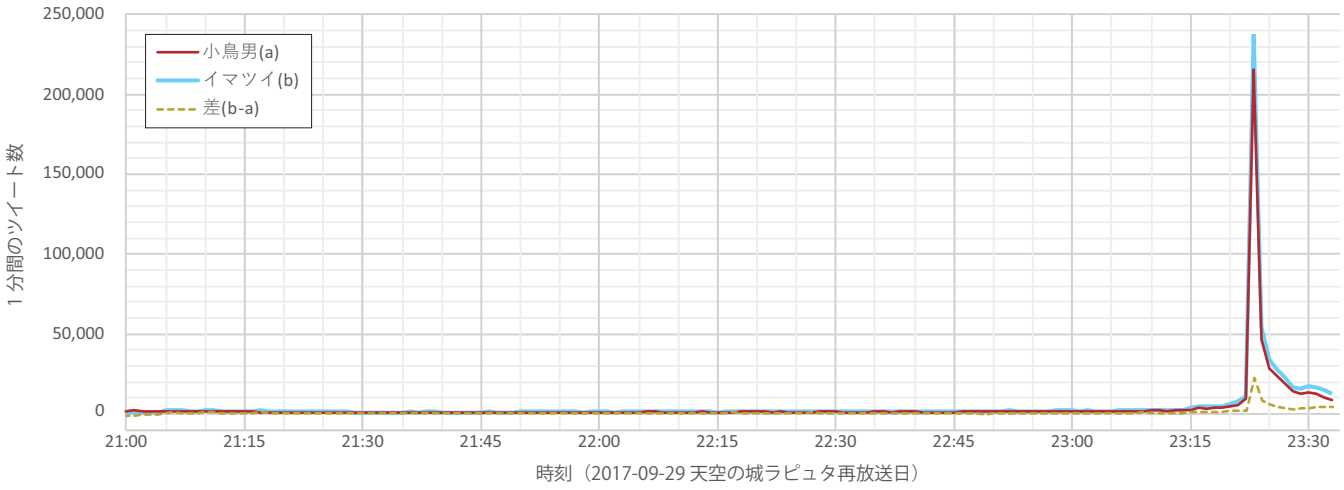

図 3　2017-09-29 天空の城ラピュタ (再放送) で観測された「バルス」等を含むツイート数

その結果，番組放送中に観測されたバルスを含むツイートの 1 分あたりの数は，図 3 のように推移していたことがわかった．なお，本研究で収集された数と，NTT データが公開している数，そして両者の差を併記してある．

当日の「バルス」ツイート大量発生のピークは 23 時 23 分であったが，この 1 分間の収量は，小鳥男が 21 万 5245 個，NTT データが 23 万 7295 個で両者の差は 10%未満であった．また，放送中の全収集ツイート数は同様に，59 万 2382，70 万 4893 で約 16%の差であった．両者の差の原因は，主に検索クエリによるものではないかと考えられるが，NTT データ側のクエリは公開されていないため断定はできない．

5.2　「けものフレンズ」ツイート

もう一つは，アニメ「けものフレンズ」やそれに関連した主要なキーワードを含むツイートの収集による検証である．

「けものフレンズ」[12] は，2017 年の 1〜3 月にテレビ放送されて社会現象と呼ばれるまでに人気を博したアニメ作品であり，Twitter トレンド大賞 2017 のアニメ部門で大賞にも選ばれた．この賞は，Twitter トレンド大賞実行委員会が Twitter JAPAN 協力のもとに，日本国内 4500 万人の Twitter ユーザが実際に投稿したツイートを分析することで映画・ゲーム・アニメ・ニュース等の部門別に最も流行した語を見つけ出し，授与したものである．公式ページには「4500 万人がつくった忖度なしの今年のトレンド」という説明があることから，授賞対象は人為的に選んだのではなく，機械的な分析結果に基づいて選定されたものと思われる．

さらに，NTT データ「イマツイ」によれば，けものフレンズに関するツイートは，2017 年に新語・流行語大賞にノミネートされた言葉の中で最も多いツイート数を発生させた語で，その数は年間約 5000 万であった [13]．

前項のバルスが瞬間的なコンテンツであったのに対し，けものフレンズは年間を通して話題が持続したコンテンツであり，かつツイート数に関する賞を受賞したものである．本提案手法によるこのコンテンツの収集・分析の実用性が示せれば，大多数のコンテンツに通用すると言えるため，その意義は大きい．

5.2.1　ツイートデータ収集

主要なけものフレンズ関連語（後述）を含むツイートの収集は，2017 年 3 月 20 日から 2018 年 3 月 19 日までの一年間実施した．この開始日は，アニメ「けものフレンズ」の全 12 話中第 10 話まで放送された後であるが，本提案手法の検証に利用しようと起草したのがこの頃であったため，第 1 話からの収集には間に合わなかった．

また，Twitter トレンド大賞 2017 を受賞したことや，年間約 5000 万ツイートを発生させたという事実も，収集開始後に知ったことである．本提案手法は，7 日前より古いツイートの収集ができないという制約を解決するものではないので，ツイート数に関する賞を獲得すること（本提案手法の実用性を示せること）がわかっていて収集し始めたわけではないことに注意されたい．

そして，この検証で使用したコンピュータの主なスペックは次のとおりである．

■ 2017 年 3 月から 12 月
CPU:Intel Core 2 Duo E8400，メモリ:4GB，HDD:1TB，OS:CentOS 5.11，インターネット回線:100Mbps

■ 2018 年 1 月から
CPU:Intel Core i5-6500, メモリ:4GB, HDD:3TB, OS:FreeBSD 11.1，インターネット回線:100Mbps

長期間収集作業を行う事情により，バルスの検証時とはまた別のコンピュータを使用した．また，OS が古くなっていたため，途中から別のコンピュータにデータ一式を移して収集・分析作業を継続させた．

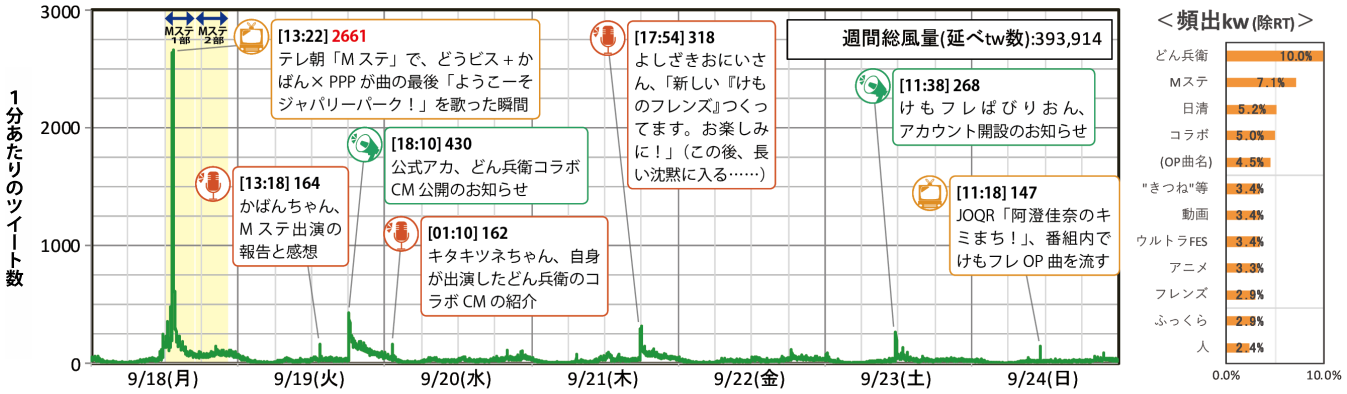

図 5　2017-09-18 週の毎分のツイート数推移と頻出キーワード

ツイート収集の際に使用した検索クエリは次のとおりである．

```
けものフレンズ OR けもフレ OR けもふれ OR
ケモフレ OR (kemono friends) OR
じゃぱりぱーく OR ジャパリパーク OR
"サーバルちゃん" OR "かばんちゃん" OR
"ペパプ"
```

時が経つにつれて関連語は増えていくが，今回の検証で上記のクエリで固定した．

5.2.2　分析

バルスツイート分析の際に行った毎分のツイート数を数える分析も含め，主に次のような分析を行った．

- 毎分のツイート数の推移をグラフ化し，極大点がどの日時に生じたか，およびどんな話題が原因で極大になったのかを調査
- 同様のことを毎秒のツイート数で実施（主に 30 分アニメ放送中に，どのシーンで極大が生じたかを秒単位で特定する場合）
- 一週間分の全ツイート本文を形態素解析して頻出語のランキングを作成し，その週で話題だったキーワードを調査
- 同様のことを番組放送時間帯で実施（番組中に最も話題になった登場人物の調査等のため）
- 一部のツイートに含まれる緯度・経度情報を地図にマッピングし，多数ツイートが投稿された場所を調査

既に述べたとおり，ツイートデータは UNIX コマンドで扱い易いプレーンテキスト形式で格納されているため，これらの分析でもその大半は，POSIX 規格で用意されている UNIX コマンドや簡単なシェルスクリプトで実現できた．

例えば，ツイート数推移グラフのある極大点について，どんな話題がその原因になっていたのかを調べるには，図 4 のようなコマンドを実行すればよい．グラフを作図して

```
cd /PATH/TO/TWEET_DATA/DIR
cat YYYYMMDD/hh/hhmm* |
awk '{print $6;}'      |
sort                   |
uniq -c                |
sort -k 1nr,1          |
head
```

図 4　極大点の原因になったリツイート文を調べるコマンド

明らかになった極大日時 “*YYYY-MM-DD hh:mm*” に該当するファイルを一通り開き，awk コマンドでツイート本文の文字列を取り出し，あとは sort と uniq -c の組み合わせで同一文字列の出現回数の多いものを列挙すれば，どんなツイートがその瞬間にリツイートされて極大を生じさせたかがわかる．一方，テレビで盛り上がるシーンが放送されるなどして，リツーイトではないツイートがたくさん投稿されて極大が生じる場合には，単純に less コマンド等で該当日時のツイートを閲覧すれば大抵は理由がわかる．

POSIX コマンドではできない作業は，グラフや地図の作図と形態素解析であった．作図に関しては表計算ソフト（今回は Microsoft Excel を用いた）や無料の Web マッピングサービス，形態素解析には mecab コマンドを用いた．ただし頻出単語の数を数える場合には，mecab コマンドで単語に分解した後，やはり先程と同様に sort と uniq -c の組み合わせで行える．

次に図 5 は，実際の分析例を示す．

これは 2017 年 9 月 18〜24 日の一週間分における「けものフレンズ」関連ツイートを分析したものである．二つのグラフから構成されており，左が分間ツイート数の推移，右が頻出語を出現頻度（登場した延べ回数を全ツイート数で割ったもの）順に並べた棒グラフである．

左の折れ線グラフでは，9 月 18 日に顕著な極大点が出ており，この時刻のツイートを調べた結果，音楽番組のミュージックステーション（M ステ）にけものフレンズの出演声優が登場して歌唱している時間帯だった．けものフレンズの人気が本放送終了半年後でも依然高いことやテレ

ビが瞬間的な話題を起こしやすい性質のメディアであることを示唆する結果が得られた．

ところが右の棒グラフを見ると，この一週間の間，「Mステ」（「ミュージックステーション」等の同義語と合算）よりも「どん兵衛」というキーワードの方が話題にされた機会が多かったことがわかる．これは翌 19 日にインターネット上でけものフレンズと日清どん兵衛がコラボした CM 動画作品が発表された影響であった．この動画はいつでも視聴可能な状態で公開されたが，インターネットがテレビとは対照的に，瞬間的ではないが継続的に話題が広まりやすい性質のメディアである様子が読み取れる．

この他にも，位置情報付のけものフレンズ関連ツイートの緯度・経度をマッピングしたところ，各地の動物園から投稿されているものが多いなど，興味深い結果が色々明らかになった [14][15].

6. 提案手法によるツイート収集・分析の実用性

今回実施した，実際の社会現象ツイートの収集・分析検証実験から，提案手法がどの程度有効であるか確認することができた．

まず，バルスツイートによる検証であるが，一つ重要な知見が得られた．それは，バルスのように瞬間的に大量のツイートを発生させるイベントであっても，100 万ツイートのオーダに収まるということである．Standard search API では 1 秒あたり 50 ツイート収集可能であるため，理論上，5〜6 時間程度で収集できることになるが，実際にも，今回の 59 万 2382 ツイートの収集に要した時間は 3 時間 26 分であり，予測を裏付ける結果となった．この所要時間は，API に課せられた過去 7 日間という制限期間を遥かに下回っている．

このように，瞬間的に大量のツイートを発生するイベントに対しても，

- Standard search API が課す制限期間に対し，十分早く収集できること
- 収集されたツイート数は，信憑性の高いデータと突き合わせても 16%の差に収まったこと

が示せたことから，瞬間的に大量のツイートを発生させるイベントに対する本提案手法の有効性が実証できた．

一方，けものフレンズ関連ツイートによる検証実験からも，重要な知見が得られた．NTT データ「イマツイ」は，けものフレンズの 2017 年のツイート総数は約 5000 万と発表していたものに対し，本検証実験により収集されたツイート数は 2316 万 6588 であり，半分程度だった．しかしながら，本研究では 3 月下旬からの一年間であって収集次期が若干遅く，また検索クエリが異っていたこと（イマツイのクエリは非公表，特に本研究では検索クエリにハッシュタグを含めていなかった）ことを考慮すれば，想定される差であり，収集時期と検索クエリを揃えられればイマツイの結果にほぼ近づけられた可能性がある．

次に，本提案手法により収集し，格納されたツイートデータの総サイズは約 89GB となり，社会現象級のツイートであっても 2018 年現在の個人向けパソコンで十分取り扱える規模に収まっていることが確認された．

また，一年間毎日収集を実施してきたが，一日分を収集するのに一日以上要することは一度も起こらなかった．ツイート数が最も多かったのは 2017 年 3 月 29 日（最終話放送日）であったが，この一日分（日本時間の 2017 年 3 月 28 日午前 9 時から 2017 年 3 月 29 日午前 9 時）で収集されたツイート数は 77 万 6796 であり，4 時間 19 分で集められた．

そして，2018 年 1 月より収集するコンピュータを更新したが，更新した後もプログラムは全く問題なく動作し，収集・分析作業に支障を来すことがなかったことも重要な点である．

このように，継続的に大量のツイートを発生するイベントに対しても，

- 本提案手法によるツイート収集量を信憑性の高いデータと比較した結果，その差は想定される範囲に収まったこと
- 収集されたツイートデータのサイズは，2018 年現在，既に個人で容易に扱えるサイズに収まったこと
- 一日分を収集するのに一日以上要する事態は一度も発生しなかったこと
- 収集作業期間中に，コンピュータを更新しても，プログラムは正常動作し，作業に支障が出なかった

が示せたことから，継続的に大量のツイートを発生させるイベントに対しても本提案手法が有効性であることが実証できた．

また，この提案手法の基本方針は，ビッグデータ分野の研究を遂行する上でも応用可能である．分析に必要なデータのみを早い段階で選択・抽出することはもちろん，プログラムやデータフォーマットを平易な構造にし，分析データおよび分析作業環境の持続可能性を高めることは，同分野においても求められると考えられ，その際には本研究で得られた知見が活かせる．

7. まとめ

ツイートの分析は，多くの人々が，多くの視点から分析するほど，社会にとって興味深い結果がより多く明らかになる．よって誰でも簡単に始められるように裾野を広げるべきと筆者らは考えている．しかしながら，社会現象等の膨大な量のツイートを発生させる話題を取り扱おうとした場合，費用的，技術的な問題から，それが可能なのは十分な予算や設備を持った大企業や組織に限られていた．

そこで本研究では，個人にとっても実用的な大量ツイート収集・分析の手法を提案すると共に，提案に基づく収集

プログラムを作成した．

提案手法に，POSIX 中心主義というプログラミング指針を採り入れた理由の一つも，裾野を広げるためである．多くの環境で，簡単に，かつ長期間使える手法であれば，より多くの人々がツイート分析を始められるようになり，今後ますます興味深いツイート分析が報告されることが期待される．

謝辞

本手法を支持し，本論文投稿にあたり御指導くださった金沢大学の共同研究者の皆様，USP 研究所の皆様に，心より感謝を申し上げます．

参考文献

[1] 鳥海不二夫，Twitter 上のビッグデータ収集と分析，組織学会 組織科学 48(4)，47-59，2016-04-07.
[2] 松村 真宏，TTM: TinyTextMiner β version (オンライン)，入手先 ⟨http://mtmr.jp/ttm/⟩（参照 2018-05-12）.
[3] Twitter,Inc.，Standard search API (オンライン)，入手先 ⟨https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets⟩（参照 2018-05-12）.
[4] NTT データ，Twitter データ提供サービス サービスメニュー (オンライン)，入手先 ⟨https://nazuki-oto.com/twitter/service_menu.html⟩（参照 2018-05-12）.
[5] Twitter,Inc., Pricing – Twitter Developers (オンライン)，入手先 ⟨https://developer.twitter.com/en/pricing⟩（参照 2018-05-12）.
[6] NTT データ イマツイ，全量データで振り返る、2016 の激動！ (オンライン)，入手先 ⟨http://imatsui.com/trend/201612/⟩（参照 2018-05-12）.
[7] 松浦智之, 大野浩之, 當仲寛哲，ソフトウェアの高い互換性と長い持続性を目指す POSIX 中心主義プログラミング，デジタルプラクティス 8(4)，352-360，2017-10-15.
[8] 秘密結社シェルショッカー日本支部，恐怖！小鳥男 (オンライン)，入手先 ⟨https://github.com/ShellShoccar-jpn/kotoriotoko⟩（参照 2018-05-12）.
[9] Twitter Japan, 2013-08-03 12:01:14(JST) のツイート (オンライン)，入手先 ⟨https://twitter.com/i/web/status/363494742518013952⟩（参照 2018-05-12）.
[10] 秘密結社シェルショッカー日本支部，擬似リアルタイム Twitter 検索・収集コマンド“gathertw.sh”使い方 (オンライン)，入手先 ⟨https://github.com/ShellShoccar-jpn/kotoriotoko/blob/master/APPS/gathertw.md⟩（参照 2018-05-12）.
[11] 宮崎駿 (監督)，天空の城ラピュタ (DVD)，JAN:4959241980144
[12] けものフレンズプロジェクト，けものフレンズ (オンライン)，入手先 ⟨http://kemono-friends.jp/⟩（参照 2018-05-12）.
[13] NTT データ イマツイ，リアルな話題量ランキング『イマツイ ツイート大賞 2017』を発表！ (オンライン)，入手先 ⟨http://imatsui.com/seasonal_topics/post_143/⟩（参照 2018-05-12）.
[14] 松浦リッチ研究所，ついったーちほーの最大瞬間風速，2017 年 8 月 11 日発行.
[15] 松浦リッチ研究所，フレンズかんそくたい，2017 年 12 月 29 日発行.