4B-02

小規模構成で実施可能な大量ツイート分析手法の提案(1) 「バルス」ツイー トを対象とした収集方法検証

松浦 智之† 當仲 寛哲† 大野 浩之‡

ユニバーサル・シェル・プログラミング研究所† 金沢大学 総合メディア基盤センター‡

1. はじめに

半角 280 文字以内の文章が投稿できる SNS「Twitter」に投稿された文章（ツイート）を収集・分析し，世論調査やマーケティング等の材料に活用する事例が近年増えてきた．しかし，大手企業や団体，又は専門業者による事例はよく目にするものの，個人など小規模で実践している例はほとんど発表されていない．主な理由は，それら業務向けの Twitter API の利用料[1]や機材が個人には高額であることだと思われる．

そこで本研究では，無料で公開されている Twitter API のみの利用，かつ取得方法を工夫することで，PC1 台でのツイートの収集・分析を現実的なものとする方法を提案する．

2. 提案するツイート収集・分析手法

2.1. 基本的なアイデア

Twitter の月間アクティブユーザ数（MUA）は，2016 年の平均で 3 億 1900 万人[2]で，つまり月に最低 3 億 1900 万個（1 秒あたり 123 個）の投稿がされている．一方，無料の Twitter API の一つである Standard search API を使用した場合に取得可能なツイート数[3]は 1 秒あたりに換算すると最大 50 個であるため，全世界のツイートをリアルタイムに収集することは不可能である．

しかしながら，実際の世論調査やマーケティング等を目的とした分析では分析テーマが存在し，必要なツイートは全体のごく一部である．そこで，全ツイートを取得してから不必要なものを除外するのではなく，今の分析テーマに必要なツイートのみを適切な検索クエリによって選択的に取得し，初めから処理対象のツイートが少ない状態にしておくことで，PC1 台で分析可能な規模に抑えることを目指す．

具体的には，前述の Standard search API をレートリミット（2018 年 1 月現在，秒間 50 ツイート取得可能な頻度）内で繰り返し呼び出し，過去のツイートへ遡りながら検索・取得する．

2.1.1. 本手法の特長と制約

ツイートの大量収集で利用が検討される API には，前述の API が属する REST API というグループの他に，Streaming API というグループがある．こちらのレートリミットは REST API よりも緩やかであるが，日本語のように分かちのない文字列中に含まれる単語の検索に対応していないことやリアルタイムに投稿されたツイートしか収集できないという欠点がある．これに対し，REST API では日本語の単語も検索でき，過去 1 週間分のツイートまでは非リアルタイムに取得可能という特長がある．

ただし，これは同時に 1 週間より前のツイートは取得できないという制約でもあるため，例えば「今年の各ノーベル賞受賞者に対する昨年の話題」のように，何らかの出来事が起こった後でそこから一週間より前のツイートが必要になるような分析テーマには向かない．

2.2. 基本的な UNIX コマンドのみによる実装

前述のアイデアを元にプログラムを製作するにあたり，後述する我々の平素からの研究方針に則り，どの UNIX 系 OS にも実装されている基本的な UNIX コマンドおよびシェルのみ利用する方針とした．したがって，収集されたツイートデータもそれら UNIX コマンドで扱える平易な形式（1 ツイート 1 行のプレーンテキスト）で格納することとした．これは次の考えに基づく．

収集されたデータは長年にわたり貴重な資料になり得るものであるため，後世の研究者も簡単に参照可能な状態で保存されるべきと考える．もし，既存の収集用のプログラムを用いてデータを収集しても，それが数年でサポート終了するようなものであったとしたら，収集したデータがそれ以降活用できる保証がなくなってしまう．実際にも，既存の収集プログラムを探してみたものの，2017 年時点で公開かつサポートされているものを見つけることはできなかった．

3. 収集プログラムの作成

著者らは，前述の方針に基づいて Twitter 収集プログラム「恐怖！小鳥男」（以下、小鳥男と称す）[4]を作成した．小鳥男は POSIX 中心主義[4]と呼ばれるプログラミング指針で実装されたシェルスクリプトである．前述のとおり，プログラムは基本的な UNIX コマンドとシェルスクリプトで構成され，出力されるデータも UNIX で扱い易いテキスト形式であるため，両方とも高い可

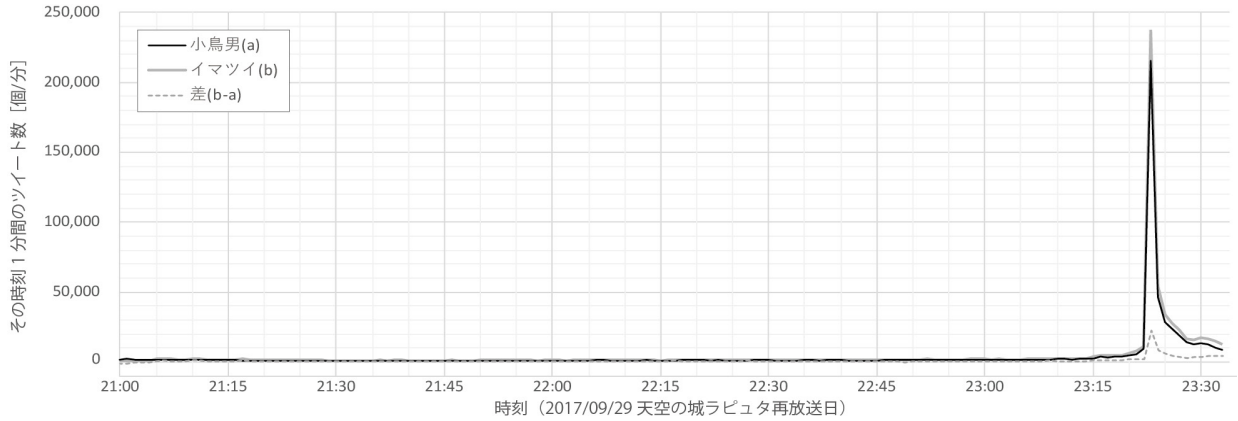

図 1. 2017/09/29 の天空の城ラピュタ再放送における「バスル」ツイート収集結果比較

搬性と持続性を有する．

4. 提案手法の有効性検証

第 2 節で，ツイートの収集段階でデータを絞り込めば個人でも分析できる可能性を指摘したが，実際に可能であるどうか，具体的な分析テーマを定めて検証を行った．

4.1. 「バルス」によるサービス停止事件

2009 年にアニメ映画「天空の城ラピュタ」がテレビで放送され，劇中で登場人物が「バルス」と叫んだ際，多くの視聴者も一斉に「バルス」を含むツイートを投稿したことで Twitter 社のシステムがサービス停止に追い込まれる事件が発生した．その後はシステムが増強され，再発しなくなったものの，大量ツイートが投稿されるイベントとして象徴的なものになった．

4.2. 「バルス」ツイートの収量比較

2017 年 9 月 29 日にも同作品が再放送され，大量の「バルス」ツイートが発生したため，小鳥男を用いてそれらを可能な限り全て収集し，ツイート数を数える試みを行った．その際の検索クエリは，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

同じ試みは NTT データのニュースサイト「イマツイ」が毎回実施し，結果を公開しているため，その収量との比較を行った（ただしイマツイは詳細な検索クエリを公表していない）．

なお NTT データは，Twitter 社のデータを公式に代理提供する企業であり，収集されたツイート数の信憑性は高いと思われる．

4.3. 収集結果

当日の放送時間帯（21:00～23:34）の各時刻の収集結果を**図 1** に示す．当日の「バルス」ツイート大量発生のピークは 23:23 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%未満となった．また，放送中の全収集ツイートは同様に，59 万 2382 個，70 万 4893 個で約 16%の差となった．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．

この結果からわかるもう一つ重要な事実は，「バルス」のような大量ツイートが発生する事案でも， 100 万ツイート程度の規模だということである．前述の Standard search API では 1 秒あたり 50 個収集可能であるから，仮にその 10 倍のツイート数になる分析テーマがあったとしても 2 日程度で集められる計算になり，1 週間以内という制約下でも十分に収集できる．

5. まとめ

本研究では，収集方法を工夫することで個人でもツイート収集・分析を可能にする方法を提案した．しかし，バルスの例は短期的な事案に過ぎず，十分な例とは言い難い．そこで，より長期的な分析例について，第 2 報[6]にて報告する．

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に，心より感謝を申し上げます．

参考文献

[1] NTT データ. Twitter データ提供サービス. https://nazuki-oto.com/twitter/ (2018 年 1 月 8 日閲覧).

[2] EDGAR Online, Inc. Twitter Inc. FORM 10-K (Annual Report) http://files.shareholder.com/downloads/AMDA-2F526X/4222743329x0xS1564590-17-2584/1418091/filing.pdf (2018 年 1 月 8 日閲覧).

[3] Twitter, Inc. Standard search API — Twitter Developers. https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets (2018 年 1 月 8 日閲覧).

[4] 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. 情報処理学会デジタルプラクティス,8(4),352-360

[5] NTT データ イマツイ.『バルス祭り』、今年はリアルタイムで盛り上がりを中継！ http://imatsui.com/seasonal_topics/post_136/ (2018 年 1 月 8 日閲覧).

[6] 松浦 智之, 當仲 寛哲, 大野 浩之. 小規模構成で実施可能な大量ツイート分析手法の提案(2) ―『けものフレンズ』ツイートを対象とした分析方法検証, 情報処理学会第 80 回全国大会, 2018.