

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

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

本研究においては第1報^[1]にて、まずSNSの一つ「Twitter」に投稿された文章（ツイート）の収集・分析が、個人やそれに類する規模の組織にとっては現状では費用が高すぎる等の理由により実現が難しい点を指摘した。そして、そのような個人であっても、収集・分析を実現する方法を提案し、収集事例を報告した。しかしながら、そこで報告した事例は短期的な例であり、提案手法の実用性を示すには十分とは言いが難かった。

そこで第2報では、一つの分析テーマを約9か月間、収集・分析した例を報告する。

2. 分析対象としたツイートについて

2.1. けものフレンズ

本研究では「けものフレンズ」とそれに関連する関連語のいくつかをツイートを分析テーマとして選び、2017年の3月20日から12月17日までの約9か月間、収集・分析を行った。「けものフレンズ」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、後述のTwitterトレンド大賞2017のアニメ部門で大賞にも選ばれた。

ただし第1報で説明したように、本研究の提案手法では1週間より前のツイートは収集できないため、受賞が決まってから収集したものではないことに注意されたい。同年3月中旬にこの作品が著しく話題になっていることに注目して収集を開始したものである。

2.2. Twitterトレンド大賞2017

Twitterトレンド大賞2017は、Twitterトレンド大賞実行委員会がTwitter JAPAN 協力のもと授与した賞である。日本国内4500万人のTwitterユーザーが実際に投稿したツイートを分析し、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、2017年12月23日に授与対象の用語を発表した。「4500万人がつくった付度なしの今年のトレンド」という説明^[2]から、授与対象は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。

3. ツイートの収集と分析方法

3.1. ツイートの収集方法について

けものフレンズといくつかの関連語のツイートは、第1報で示したプログラム「恐怖!小鳥男」(以下、小鳥男と称す)^[3]にて収集した。その際の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"

3.2. データの分析方法について

今回は、主に次の二つの観点に基づくツイート分析を行った。

3.2.1. 各時刻1分間のツイート数集計

一つは、各時刻の1分間のツイート数を数え、その推移に基づくトレンド分析である。1分間あたりのツイート数の推移を折れ線グラフ化すると、極大になる日時が多数観測される。それぞれの日時では、多くのユーザーが何らかの共通の出来事を話題にしていることが多い。そこで、グラフを描画して極大をとる日時を明らかにし、その日時のツイートが格納されたファイルを開覧すれば具体的に何が起こっていたのかわかる。

第1報のとおり、収集されたデータは1ツイート1行のテキスト形式で格納されており、かつ、ツイートが投稿された年月日時分秒ごとにファイルとディレクトリが分かれているため^[3]、各ファイルの行数を数えれば各時刻の秒間ツイート数がわかる。これを分単位に集約すれば、日時対分間ツイート数も容易に求められる。尚、これらの操作は、基本的なUNIXコマンドであるwc, find, awk, sedなどを組み合わせるだけで実現できる。さらに、sedやtrコマンドを用いてCSVファイル化すれば、各種表計算ソフトで開けるので、グラフ化まで簡単に行える。

3.2.2. 形態素解析による頻出語調査

もう一つは、取得したツイートを形態素解析器に掛けて単語毎に分解し、各単語の出現回数を数えて頻出語を調べることによるトレンド分析である。前述のグラフの極大時刻に基づくトレンド分析では瞬間的に盛り上がった話題が見つかりやすいのに対し、こちらの方法では静か

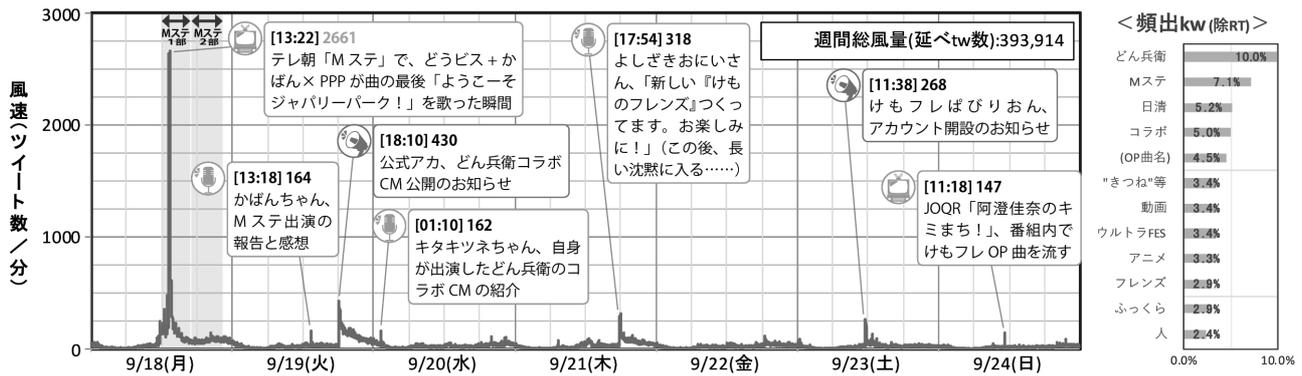


図 1. 2017/09/18-24 のけものフレンズ関連ツイート分析例 (分間ツイート数推移グラフと頻出語グラフ)

ではあるが継続的であった話題を見つけやすい。

今回、形態素解析器には mecab コマンドを用い、分解された語の勘定には sort, uniq などの基本的な UNIX コマンドを用いた。

4. 分析結果

4.1. 分析例 (9月18日の週)

分析結果として作図したグラフの一例を図 1 に示す。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の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。

4.2. 分析の総括

この 9 か月間、他にも興味深い様々な分析結果が得られた。その結果は、文献^{[4][5]}にまとめた。また、対象となるツイートが提案手法の限界を超えるほどに大量に発生し、収集しきれなくなる事態はこの期間には一度も発生しなかった。

そして、この期間に 3.1 項のクエリで取得できたツイートは 2070 万 9555 個であった。

一方、Twitter 社のデータを公式に代理提供する企業である NTT データのニュースサイト「イマツイ」の報告によると 2017 年の「けものフレンズ」のツイート総数は約 5000 万であったという^[6]。イマツイと本研究で使った検索クエリは恐らく異なるために単純な比較はできないが、本研究では 3 月下旬から収集を開始したことや、ハッシュタグを含めていなかった分収集できたツイート数が少なかったであろうことを鑑みれば、本提案手法によるツイート収集・分析はある程度の実用性があると言える。

5. まとめ

ツイートの収集・分析自体は既に珍しい技術ではないが、現時点で同等のことを個人で行うには費用の負担が大きい。しかし、収集や分析方法を工夫をすることにより、無料の Twitter API と PC1 台の小規模な構成でも実用的なツイート収集・分析が実現できることを示した。

引き続き、さまざまな突発事象に臨機応変に対応して分析を行う予定である。

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] 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(1) — 『パルス』ツイートを対象とした収集方法検証, 情報処理学会第 80 回全国大会, 2018.
- [2] トrend大賞@リプレイ配信中 みなさまありがとうございます! <https://twitter.com/TrendAward> (2018 年 1 月 8 日閲覧).
- [3] 秘密結社シェルショッカー日本支部. 恐怖! 小鳥男. <https://github.com/ShellShoccar-jpn/kotoriotoko/> (2018 年 1 月 8 日閲覧).
- [4] 松浦リッチ研究所 編集・発行. ついつつ一ちほーの最大瞬間風速. 2017 年 8 月 11 日発行.
- [5] 松浦リッチ研究所 編集・発行. フレンズかんそくたい. 2017 年 12 月 29 日発行.
- [6] 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! http://imatsui.com/seasonal_topics/post_143/ (2018 年 1 月 8 日閲覧).