Twitterトレンドのリツイートによる分析と時系列の可視化

Analysis of Twitter Trends by Using Retweets and Visualization of Their Time Series

小西　敦郎
Atsuro Konishi
法政大学情報科学部コンピュータ科学科
E-mail: atsuro.konishi.6b@stu.hosei.ac.jp

Abstract

Twitter trends are a function that extracts and presents keywords and hashtags that become a popular topic in Twitter. Typically, one trend word or hashtag is related to thousands of tweets. It is difficult to understand all of these tweets in a short time by using the standard display method and sort algorithms provided by Twitter. Most of previous studies analyze and visualize tweets by using text-based classification methods. However, these suffer from the accuracy of classification results, because a typical tweet has only poor textual information. This paper presents a system that analyzes tweets related to Twitter trends by combining a retweet clustering method and a time-series visualization to allow users to understand a topic flow of Twitter trends. This system also provides effective legends and a display of practical tweets with images, because the clustering result has little information from which users can understand its contents. Legends are extracted from text of tweets in clusters by using feature words extracting methods, tf-idf, BM25, or word frequency. Actual tweets to display are prioritized by non-textual data like a reverse chronological order, and numbers of favorites or images.

1. はじめに

Twitter トレンド(以下トレンドと呼称)とは，Twitter 上で話題になっているキーワードやハッシュタグをリアルタイムに抽出する Twitter の機能である．Twitter では各トレンドに関連するツイートを取得することができ，1 トレンド当たり 1000 件を超える関連ツイートが存在する．Twitter の標準ソート法(話題，最新など)や縦 1 列に並べる表示法では，大量にある関連ツイート全体の意見の推移を短時間で把握することは難しい．また，過去の研究ではツイートのテキストをもとに分析・可視化をしているため，精度に不安がある．

本研究ではリツイートによるツイートの分類と時系列に沿った可視化を組み合わせ，短時間でのトレンドの全体像の理解を容易にする．分類に文字情報を用いないため，可視化結果は内容を表す情報に乏しい．理解を補助するために，特有な単語の抽出や画像や URL を含むツイートの表示を行う．単語の抽出法やツイートを表示する順番は複数用意し，対象とするトレンドや要求に合わせて切り替えることができるようにする．

2. 関連研究

Twitter のトレンドに関する研究はあまりないが，トレンド機能のアルゴリズムの信頼性を評価するもの [1]や，Twitter トレンドをリアルタイムで 4 つのテーマ(ニュース，開催中のイベント，ミーム，記念事)に分類するもの [2]がある．後者は教師あり学習のサポートベクターマシンを利用しており，テーマの社会的特徴と文書中の単語出現数をもとに分類している．

ツイートを分類して可視化する研究は多く行われている．Wenwen ら [3]はトピックモデルである LDA によりツイートを分類し，ThemeRiver により話題の推移を可視化している．LDA は，文書にはトピックが存在すると仮定し，教師なし学習により出現する単語の生成確率をもとに文書をトピックごとに分類する手法である．Florence [4]らの SentiCompass はツイートのテキストを感情により分類し，可視化している．可視化には，1 つの時間帯における分類結果を環状のヒストグラムとし，時系列となるよう同心円状に配置する手法を用いている．

3. 準備

本研究では，ツイートの分析に Uchida [5]のリツイートクラスタリング手法を用いる．この手法は，同じユーザにリツイートされたツイートは似たような内容であるという前提のもとツイート間の類似度を決定する．クラスタリングには，以下の 3 つの段階を踏む．

3.1. ツイート間の類似度計算

リツイートしたユーザの重複度によって，ツイート間の類似度を計算する．

ツイートiに対して，$r_{i,j}$を以下のように定義する．

$$r_{i,j} = \begin{cases} 1\ (\text{ユーザ}j\text{がツイート}i\text{をリツイートしている}) \\ 0\ (\text{ユーザ}j\text{がツイート}i\text{をリツイートしていない}) \end{cases}$$

ツイートiのリツイート関係をベクトル$\boldsymbol{t_i}$として，すべてのユーザUに対して以下のように定義する．

$$\boldsymbol{t_i} = (r_{i,0}, r_{i,1}, \ldots r_{i,U})$$

ツイートiとツイートjの類似度を，シンプソン係数を用いて以下のように求める．

$$\mathrm{sim}(\boldsymbol{t_i}, \boldsymbol{t_j}) = \frac{|\boldsymbol{t_i} \wedge \boldsymbol{t_j}|}{\min(|\boldsymbol{t_i}|, |\boldsymbol{t_j}|)}$$

3.2. 重み付き無向グラフの構築

全てのツイートの組に対して 3.1 節の計算を行う．求めた類似度の上位 N位までのツイート組をリンクさせ，重み付き無向グラフを構築する．

Supervisor: Prof. Hiroshi Hosobe

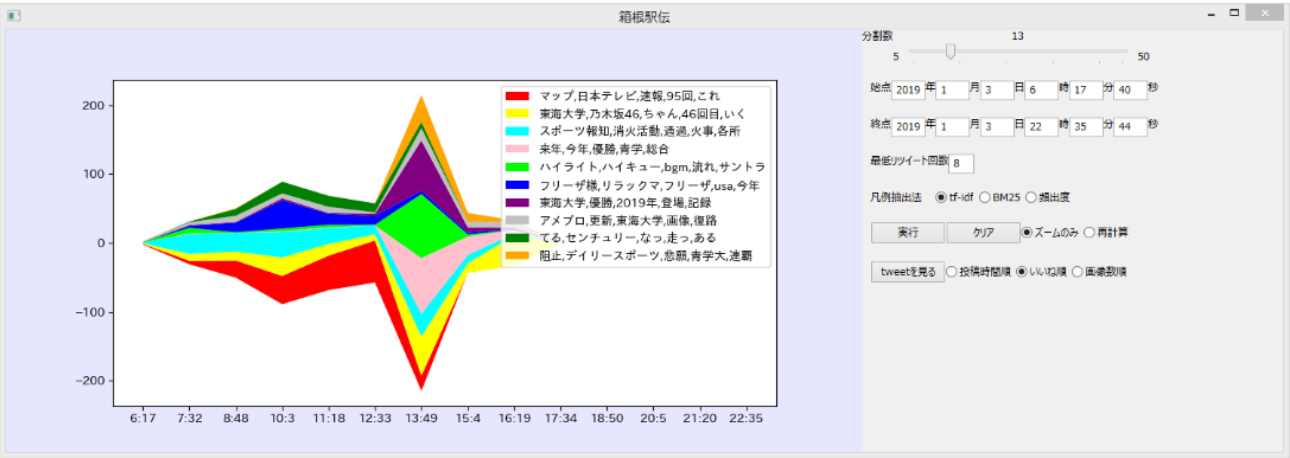

図 1　トレンド“箱根駅伝”に関連するツイートのクラスタリング結果の ThemeRiver を用いた可視化

3.3. Louvain 法によるクラスタリング

3.2 節で構築した重み付き無向グラフに対して Louvain 法 [6]を用いることでクラスタリングを行う．Louvain 法は，グラフの質を表すモジュール性を最大化によるクラスタリング手法である．同じクラスタに所属するツイート間の重みが大きく，異なるクラスタに所属するツイート間の重みが小さくなるようなクラスタリング結果を得ることができる．

4. 提案手法

本研究では，Uchida のリツイートクラスタリング手法と Harve の ThemeRiver [7]を用いた可視化を組み合わせ，ツイートの分析を行う．分析対象となるツイートはトレンドのキーワードを Twitter で検索して得たものであり，トレンド関連ツイートと呼称する．

可視化自体は各話題の内容を理解するための情報に乏しい．そのため，凡例を抽出と実際のツイートの表示により理解を補助する．凡例は，各話題のツイート内容を形態素解析し，特徴となる単語を凡例として抽出する．ツイートの表示は，話題ごとに関連するツイートを画像とともに行う．

4.1. リツイートクラスタリング

リツイートクラスタリング手法の大部分は 3 章で説明したとおりであるが，3.2 節に変更点がある．すべてのツイートでなく，ユーザが定めた最低リツイート数以上で，その範囲内のツイート数が 1500 件を超えるツイートを対象とする．これは 3 章全体の計算量を抑えるためである．リツイートをリツイートしたものに関してはリツイート元のツイートをリツイートしたとして扱う．

4.2. ThemeRiver を用いた可視化

図 1 は ThemeRiver を用いた可視化結果である．ThemeRiver は話題の推移を時系列に沿って，川の流れのように可視化する手法である．流れの各色がトレンド内の話題に相当し，積み上げグラフのような形で可視化する．縦軸にツイートの強度，横軸に時間をとる．4.1 節でのクラスタリング結果を，ThemeRiver での話題として扱う．ある期間におけるツイート数をカウントし，ThemeRiver の縦軸方向の値とする．

4.3. 凡例の抽出

各話題の凡例は，ツイート内容の形態素解析と情報検索により抽出する．凡例の抽出法として，文書中の単語の頻出度による方法と，単語の重要度による tf-idf 法，BM25 を用いる．ここでの文章は，ThemeRiver の 1 つの話題に含まれるすべてのツイート内容とする．これらの抽出法は対象とする文書によって結果の精度が異なるため，選択して切り替えることができるようにする．図 1 の凡例は ThemeRiver の色と凡例左部の色が対応している．

4.4. ツイートの表示

凡例で内容が理解できなかった場合や，実際のツイートを見たい際にツイートの表示を行う．話題ごとに，関連するツイートを画像とともに投稿時間，いいね数，画像数順のいずれかで表示する．

5. 実装

実装には Python を用いた．Twitter からのツイートの取得にはライブラリ Tweepy を使用し，グラフの描画には matplotlib，GUI には wxPython を使用した．凡例の抽出のため，形態素解析器として MeCab を利用した．トレンドに関連するツイートはトレンドとなった単語やハッシュタグでキーワード検索することで得る．

GUI 部分では，時系列の始点，終点，分割数と最低リツイート数，凡例の抽出法変更する操作が可能である．実行ボタンの右側のラジオボタンにより，前述の変更点をグラフに反映する方法を変更できる．ズームのみの場合は，クラスタリング結果はそのままに変更したパラメータを反映させて描画する．再計算の場合は，変更したパラメータをもとにクラスタリングし結果を描画する．Tweet を見るボタンにより，投稿日時，いいね数，画像数でソートしたツイートを表示する．

6. データセット

実験に用いるデータセットは表 1 である．ツイート数は，取得したツイートのうちリツイート数が 1 以上のものをカウントしている．このトレンドは，2019 年 1 月 2 日，3 日に行われた第 95 回東京箱根間往復大学駅伝競走の略称である．2 日が往路，3 日が復路であり，往復の合計タイムをもとに順位が決定する．

表 1　データセット

トレンド名	ツイート数	リツイート数
箱根駅伝	34,894	10,027,794

7. 実験

本研究の手法を用いて，データセットのトレンド関連ツイートを分類，可視化する．対象としたトレンドは 1 月 3 日のトレンド“箱根駅伝”で結果が図 1 である． 3 日は午前 8 時にスタートし 5，6 時間をかけて箱根東京間を走るという内容であり，その様子がテレビ中継されていた．図 1 におけるクラスタリング対象となったトレンド関連ツイートは表 2 の条件を満たしたツイート 1580 件であり，凡例は tf-idf 法で抽出した．グラフより，ゴールのあった時間帯である 13 時 49 分ごろにツイート数が急増していることがわかる．また，ゴールの時間帯と競技途中の時間帯では出現する話題が異なり，それぞれの時間帯に対応した話題となっている．

表 2　図 1 の分析対象となったツイートの条件

最小リツイート数	8
最大リツイート数	14
期間始点	1 月 3 日 6 時 17 分 40 秒
期間終点	1 月 3 日 16 時 35 分 44 秒

競技途中の話題について詳しく説明する．赤色の話題は，日本テレビが運営する箱根駅伝の速報を知らせる web サイトについてである．後述する競技途中の他の話題が競技内容に直接かかわるものでないため，主にこの web サイトを用いて競技内容の確認や伝達をしていたと考えられる．10 時 3 分ごろに全体のツイート数が増加しているが，これは主に青色で示した話題に含まれるツイートが増加したことによるものである．この話題の凡例には，“フリーザ”，“リラックマ”と駅伝とは関係のないキャラクター名が並び，内容を把握できない．そのため，この話題について GUI を用いて実際のツイートを表示する．

図 2 は実際に表示したツイートである．実際のツイートと画像を見ると，沿道の応援者の中に当該のキャラクターがいることがわかる．この 2 つのキャラクターについてのツイート内容を見ると，文字情報の共通点は非常に少ない．このような内容のツイートに関しても，“沿道の特徴的な応援者”に興味を持つユーザがいたため同じ話題として分類することができた．この例は本研究の手法により取得できる特徴的なものである．

図 2　図 1 の青色で示した話題に含まれるツイート

次に，ゴールの時間帯の話題について詳しく説明する．ゴールの時間帯は 13 時から 14 時の間であるため，この期間に対してズームと分割数の変更を行った．結果は図 3 であり，各話題の色は図 1 と一致している．

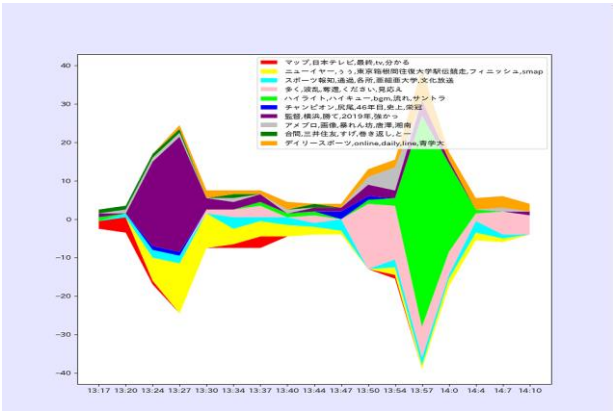

図 3　図 1 を 13 時から 14 時の期間でズームした結果

図 3 では紫色，黄緑色，ピンク色の話題に含まれるツイートがゴールの時間帯に急増している．これらの話題は凡例より，東海大学が優勝した話題，ハイライトの話題，来年や今年といった話題という事が読み取れる．図 3 のようにズームを行うことにより，これらの話題にも流れがあることがわかる．紫色の話題は 13 時 27 分に最もツイートされているが，これは復路の 1 位，2 位のチームがゴールした時間と重なる．往路の結果とこの時点での 1 位，2 位の結果から優勝が決定したためこの時間にツイート数が増加している．ピンク色の話題は，実際のツイートを見ると今回の箱根駅伝の感想や来年への期待といった内容である．この話題のツイートが増加した 13 時 50 分は最後のチームがゴールした時間 13 時 48 分の直後であり，その後往路復路の全結果をもとに黄緑色の話題

であるハイライトが作られ放映されたという話題の推移を可視化できている．

8. 議論

実験により，トレンドに関連するツイートの話題の推移を視覚的に表せることを示した．リツイートをもとに分類をしているので，実験での青色の話題ように文字情報の類似度が低いツイートを同一の話題と分類することができる．図 3 でのズームのように，もともとは同一の時間内で増加したと読み取れるツイートも，ユーザの操作により時系列による話題の推移を発見することもできる．赤色や青色の話題のように凡例から内容を理解することが難しいものについても，実際のツイートを見ることによって内容を短時間で理解することができる．

一方で，水色の話題は本研究手法で内容を理解するのに時間がかかる例である．水色の話題は凡例から“火事”，“消火活動”と読み取ることができ，ツイートのおよそ半分は競技中に発生した沿道火災についての内容であった．しかし，この話題について実際のツイートを表示すると，競技中の応援や感想といった内容のツイートが上位に表示される．そのため，他の話題より多くのツイートを閲覧しないとこの話題の主となる内容を詳しく理解することができない．この問題に対しては，凡例として抽出した単語を含むツイートを優先して表示する等，実際のツイートの表示順の種類を増やす必要がある．

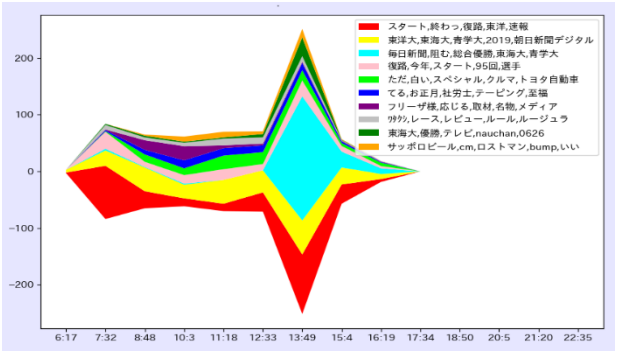

図 4 最低リツイート数を 20 とした際の可視化結果

同じトレンド関連ツイートに対して条件を変えることで，異なる結果を得ることができる．図 4 は表 2 の条件のうち，最低リツイート数を 20，最大リツイート数を 55 に変更し分析をした結果である．グラフの概形は図 1 に類似しているが，話題の内容が異なる．図 4 の水色の話題は図 1 の紫色の話題と類似しているが，実際のツイートを見ると図 1 では個人の優勝を賞賛するツイートが多いのに対して，図 4 では毎日新聞の web ニュースを引用したツイートが多い．また，オレンジ色のサッポロビールの CM に関する話題など，図 1 の結果にはない話題も出現している．最低リツイート数を変更することで分類される話題も異なるが，ユーザの手によりトレンドごとの最適な値を探すのは手間がかかる．そのため，クラスタリング結果のモジュール性などにより最適な値を推薦する機能が必要である．

4.1 節にて述べたが，計算量を抑えるためにデータセットの全ツイートうち，リツイート数により対象とするツイートを 1500 件ほどに制限している．3.2 節の計算時間はツイート数をnとすると，$O(n^2)$となる．本研究手法は GUI によりユーザが興味のある話題や期間に対して操作を繰り返すことを前提にしているため，計算精度より実行時間を優先している．しかし，特に大量のツイートを含む期間を分析する際に，ツイート数を増やすことでより精度の高い結果を得ることが期待できる．そのため，計算の高速化やデータの取捨選択といった改善点がある．

9. おわりに

トレンドに関連するツイートの話題の流れを短時間で把握するために，リツイートクラスタリングと時系列の可視化を組み合わせた．分類と可視化により，トレンドに関連するツイートの話題の推移についてグラフで表示することができた．凡例や実際のツイートの表示により短時間でのトレンド理解を補助することができた．分類の最適な条件の推薦システムの実装や凡例抽出法，ツイートの表示法の改良が今後の課題である．

文　献

[1] G. Tarleton, "Can an Algorithm Be Wrong? Twitter Trends, the Specter of Censorship, and Our Faith in the Algorithms around Us," 19 10 2011. [Online]. Available: https://socialmediacollective.org/2011/10/19/can-an-algorithm-be-wrong/.

[2] Z. Arkatiz, S. Damiano, M. Raquel and F. Victor, "Real-time Classification of Twitter Trends," *Journal of the Association for Information Science and Technology*, vol. 66, no. 3, pp. 462-473, 2014.

[3] D. WenWen, W. Xiaoyu, K. Thomas and R. William, "Identifying Topical Trends in Social Media with Topic Modeling," in *IEEE Workshop on Interactive Visual Text Analytics for Decision Making*, 2011.

[4] W. Y. Florence, S. Arnaud, K. Karsten, T. Masashiro and R. Mathieu, "SentiCompass: Interactive Visualization for Exploring and Comparing the Sentiments of Time-Varying Twitter Data," *Proc. IEEE PacificVis*, pp. 129-133, 2015.

[5] K. Uchida, F. Toriumi and T. Sakai, "Evaluation of Retweet Clustering Method Classification Method Using Retweets on Twitter without Text Data," *Proc. WI*, pp. 187-194, 2017.

[6] B. D. Vinent, G. Jean-Loup, L. Renaud and L. Etienne, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10008, pp. 1-12, 2008.

[7] H. Suran, H. Elizabeth, W. Paul and N. Lucy, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *IEEE Trans. Visual. Comput. Gr.*, vol. 8, no. 1, pp. 9-20, 2002.