Vol.2018-MPS-120 No.19
2018/9/26

極座標可視化手法を用いたトレンドワードのバースト検出

松井 直大[1,a)]　伏見 卓恭[1,b)]

概要：本研究では，現在話題になっているトピックの動向や関連性を視覚的にとらえるために，極座標平面上にトピックごとのトレンドワード出現回数をプロットする手法を提案する．具体的には，トレンドワードを含むツイートを用いて，トレンドワード間の類似度を定義する．極座標平面において，類似のトレンドを原点から見て同一方向にプロットする．さらに，時間軸を半径で表現することで，トレンドワードの広がっていく様子やバーストを表現した．実データを用いた評価実験の結果，類似のトピックを有するトレンドワードを同一方向にプロットできることを確認した．さらに，極座標平面上でトレンドワードが密集する部分を抽出できることも確認した．

Burst Detection of Trend Words Using Polar Coordinate Visualization Method

Matsui Naohiro[1,a)]　Fushimi Takayasu[1,b)]

1. はじめに

現在 Twitter のタイムライン上を流れているツイートの中から，使用頻度が高く，短時間で急上昇した話題性の高いキーワードをリアルタイムに抽出して表示してくれる機能を Twitter トレンドという．トレンドワードは Twitter 社のアルゴリズムによって決定され 5 分に 1 回の頻度で更新されるので，Twitter 利用ユーザの興味・関心の高い話題性の高いトピックをリアルタイムに知ることができる．すなわち，今起こっている話題性の高いことを知りたい場合はネットニュースやテレビ，新聞を見るより遥かに早く情報を得ることができると考えられる．しかしながら，公式 Twitter が表示しているトレンドワードはトピックの関連性や動向が表示されておらず，話題になっているトピックの中から高頻出な 1 単語または 1 文を設定している．よって，トレンドワードを見た際にどういった使い方や意味を持っているワードであるのかを調べる手間がかかる．そこで本研究では，トレンドワードを可視化しバーストを検出することによりわかりづらいトレンドワードを一目で分かるようにすることを目指す．

[1] 東京工科大学　コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology
[a)] c011530091@edu.teu.ac.jp
[b)] fushimity@stf.teu.ac.jp

2. 関連研究

Ishikawa らによる T-Scroll では，文書間の類似度計算し，k-means 法によりクラスタリングことで関連する文書群をまとめて可視化する [1]．そして，隣接する時刻におけるクラスタ間の関連度を定め，関連度の高いクラスタ間にリンクを追加することで，各時刻におけるクラスタリング結果を時間軸上にプロットする．T-Scroll のように直交座標における横軸に時間軸をとると，関連文書が時間とともに増加する場合に対応できない．さらに，各文書をクラスタにまとめて可視化するため，文書の投稿間隔に関する情報は視覚的にわからず，バースト状況を把握することも困難である．

Krstajic らの CloudLines [2] は，特定のキーワードを含むニュース記事を点として時間軸上にプロットするもので，これらの点に対してカーネル密度推定を適用することで投稿間隔が短いバーストイベントを抽出可能である．プロットされた点群は，時間軸上で連続することで線になり，重要な事象が発生する場所で線が太く表示される．

3. 提案手法

提案手法について説明する．トレンドワードの集合を U で表わし，各トレンドワード $u \in U$ を含むツイートの集合を $V(u)$ とする．同じキーワードが異なる時間でトレンドワードとなる場合があるが，それらは別のトレンドワードとして区別する．そして，ツイート $v \in V$ に含まれる一般単語の集合を $W(v)$ とする．便宜上，トレンドワード u と共起する一般単語の集合を $W(u) = \bigcup_{v \in V(u)} W(v)$ と表記する．提案手法は，以下の手順に従って，トレンドバーストを検出する：

(1) トレンドワードとツイートに出現する単語の共起関係の強さを計算する；
(2) 類似のトレンドワードが原点から見て同一方向になるように偏角を決定し，トレンドワード入りした時刻を半径とした極座標値を求める；
(3) カーネル密度推定により，極座標平面上の密分布する部分をバーストとして検出する；

ステップ 1 では，各トレンドワード $u \in U$ に対して，共起する一般単語 $w \in W(u)$ との関係の強さ $a_{u,w}$ を TFIDF により計算する．ステップ 2 では，トレンドワード u と一般単語 w の極座標ベクトル $\mathbf{x}_u$ と $\mathbf{y}_w$ を算出する．すべてのベクトルの極座標値をランダムに初期化し，正規化することで単位円上のベクトルを構築する．トレンドワード u の座標ベクトル $\mathbf{x}_u$ は，共起する単語の座標ベクトルの重み付き和で計算する：

$$\tilde{\mathbf{x}}_u \leftarrow \sum_{w \in W(u)} a_{u,w} \mathbf{y}_w, \quad \mathbf{x}_u \leftarrow \tilde{\mathbf{x}}_u / ||\tilde{\mathbf{x}}_u||.$$

同様に，一般単語 w の座標ベクトル $\mathbf{y}_w$ は，共起するトレンドワード $u \in U(w)$ の座標ベクトルの重み付き和で計算する：

$$\tilde{\mathbf{y}}_w \leftarrow \sum_{u \in U(w)} a_{u,w} \mathbf{x}_u, \quad \mathbf{y}_w \leftarrow \tilde{\mathbf{y}}_w / ||\tilde{\mathbf{y}}_w||.$$

これらを反復し収束するまで繰り返すことで，単位円上の極座標ベクトル群 $\mathbf{X} = [\mathbf{x}_u]$ を得る．これにより，類似の単語と共起するトレンドワードは，原点から見て同一方向の座標ベクトルとなる．そして，各トレンドワードの出現した時刻により半径 r_u を決定し，座標ベクトルを $\mathbf{z}_u \leftarrow r_u \cdot \mathbf{x}_u$ とする．ステップ 3 では，$\mathbf{Z} = [\mathbf{z}_u]$ に対して，カーネル密度推定により密分布している部分を抽出する．

4. 評価実験

Twitter トレンドを 5 分毎に最大 50 件自動取得し，各トレンドワードを含む直近のツイートを最大 100 件，自動収集したものを対象とする．今回用いる対象のデータは 2018 年 7 月 31 日分，11,532 トレンドワードとする．

トレンドワードを極座標平面に可視化しバースト検出をしたものを図 1 に示す．点の色はトピックごとに分類しておりトピック数は 50 種類でトレンドワードを分類したが色の種類が不足していたため，異なるトピックでも同じ色で表しているものがある．図 1 より，背景の色はトレンドワードのバーストレベルを表しており，赤いほどバーストレベルが高く青いほど低いことを示す．同色のトレンドワードがある程度まとまっていることから，トレンドワードの関連性を反映した可視化結果になっていると言える．さらに，背景色に着目すると，トレンドワードが密集する部分は赤で高いバーストレベルを示しており，関連トピックのトレンドワードがバーストしている部分を検出できていると言える．

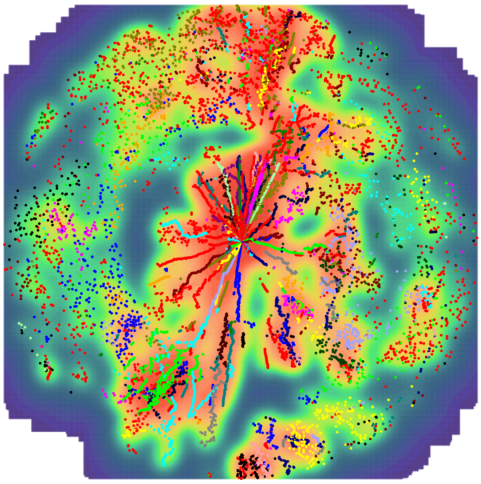

図 1　トレンドワードのバースト検出

5. おわりに

本研究ではトレンドワードを極座標平面に可視化しバースト検出を行なった．評価実験の結果，トレンドワードやトピックの動向や関連性を見ることができた．しかし，バースト検出で用いたカーネル密度推定はノンパラメトリックな手法であり，原点から見て放射方向だけでなく，円周方向に密な部分もバーストとして検出される．よって，フォン・ミーゼス分布などを仮定して，放射方向にバーストする部分のみを検出する手法を検討ことは今後の課題である．

謝辞　本研究は，JSPS 科研費（No.16K16154）の助成を受けたものである．

参考文献

[1] Ishikawa, Y. and Hasegawa, M.: *T-Scroll: Visualizing Trends in a Time-Series of Documents for Interactive User Exploration*, pp. 235–246, Springer Berlin Heidelberg (2007).
[2] Krstajic, M., Bertini, E. and Keim, D. A.: CloudLines: Compact Display of Event Episodes in Multiple Time-Series., *IEEE Trans. Vis. Comput. Graph.*, Vol. 17, No. 12, pp. 2432–2439 (2011).