外れ値による補正に基づく位置を考慮したバースト検出手法

犬塚　幹浩†　　酒井　達弘†　　田村　慶一†
†広島市立大学大学院情報科学研究科

1. はじめに

近年，Twitter 上に投稿されるツイートから，人々の関心の高い地域性を持つトピックやイベントを取り出すことが注目されている．関心の度合いを示す注目度を取り出す方法としてバースト検出が用いられており，地域性を持つトピックやイベントを取り出すにはユーザと文書データの位置を考慮してバーストを検出することが重要となる．我々は，位置を考慮した MACD ヒストグラムベースのバースト検出手法を開発しており，位置情報付きツイートに現れるトピックやイベントに関連する語句のバーストをユーザの位置に応じて抽出することに成功している．しかしながら，日常的に変動する語句の出現頻度の増減までバーストとして検出してしまうという課題が存在している．そこで本研究では，語句の出現頻度の増減を外れ値に基づき補正することでこの問題点を解決する．

2. MACD ヒストグラムベースのバースト検出手法

時間単位など一定期間でまとめた文書データ集合の系列を文書データ集合系列と呼ぶ．本研究では，ツイートを文書データ，一定期間内に投稿されるツイートを文書データ集合系列として扱う．また，語句単位でバーストを検出するものとする．ここで，特定の語句を含む文書データを関連文書データと呼ぶ．文書データ集合系列から関連文書データ数を算出し，算出した系列の MACD ヒストグラムを求める．MACD ヒストグラムは系列データの加速度を示し，正の値である場合は出現頻度が急増していると考え，当該語句はバースト状態であると判定する．次に，位置情報を考慮するために，ユーザと文書データの距離から文書データの重要度を算出する．この重要度を用いて関連文書データ数の系列を補正し，MACD ヒストグラムを求めると，位置を考慮したバースト度を算出できる．

3. 提案手法

Twitter への投稿には日常的な増減現象が存在する．例えば，地名を含むような語句は一日の中で変動があり，何もしなければ，その増減の変動をバーストとして検出してしまう．そこで，この変動の影響を無効にするために過去の関連文書データ数を統計データとして用い，外れ値に基づく補正を行う．

(A) 四分位数による補正

統計データから各語句について重要度の第一四分位数，第三四分位数を求め，四分位数を用いて補正を行う．具体的には，関連文書データ数から四分位範囲の 1.5 倍を引いた値が正値であるときだけ外れ値とみなしバーストを検知，それ以外は 0 とする．

(B) 標準偏差による補正

統計データから各語句の重要度の平均値と標準偏差を算出しておき，標準偏差を用いて補正を行う．平均値から標準偏差の 3 倍離れた値を外れ値とみなし，関連文書データ数から平均+3 σ を引いた値が正値であるときだけ外れ値とみなしバーストを検知，それ以外は 0 とする．

4. 評価実験

評価実験は，ジオタグ付きツイートのデータセットを Twitter Streaming API で取得して行った．2016 年 12 月 23 日から 2017 年 1 月 13 日に取得したツイートを統計データ，2017 年 1 月 14 日から 2017 年 1 月 16 日に取得したツイートを評価データとしてユーザが東京都庁にいることを仮定して，どのような語句がバーストとして検出されたかを補正なしと補正ありで比較する．表 1 にバースト度の累計が大きい語句の上位 3 件を示す．2017 年 1 月 14 日，15 日はセンター試験が行われた日である．補正なしでは日常からツイートされている「東京都」などのバースト度が高いため，ランキングの上位となっている．補正ありでは「センター試験」やその日に全国的に観測された「雪」などが上位となっており，その日の話題と一致しているといえる．

表 1　語句ランキング

	補正なし	四分位数による補正	標準偏差による補正
1	東京都	センター試験	センター試験
2	ジャンル	雪	雪
3	店	追加	東京ドームシティホール

5. まとめ

本研究では，外れ値に基づく位置を考慮したバースト検出手法を提案した．評価実験の結果，外れ値を用いることにより，日常的な変動を取り除くことができた．今後の課題として，イベントが行われたときそのイベント名，イベント名の略称，イベントの開催された会場などはすべて別の語句として検出されてしまうため，関連する語句をまとめる手法の検討があげられる．

謝辞

本研究の一部は，総務省 SCOPE(受付番号：162308002)と広島市立大学・特定研究費の支援により行われた．

Copyright © 2018 IEICE