4N-05

SNSアカウント分類のための社会的バースト性抽出手法

中村 聡志† 高崎 尚人‡ 小林 亜樹†

†工学院大学工学部情報通信工学科 ‡工学院大学大学院工学研究科電気・電子工学専攻

1 はじめに

Twitter では多数のユーザをフォローし、多くのトピックが混じりあってしまうと読みにくい。リスト機能はタイムラインをカテゴリ別などで分類することで読みやすくするのに使われるが、分類対象のアカウントが多くなると手間が大きい。

そこで本研究では、人手で可能な程度の規模で選別して作成されたアカウント集合に対して、任意の他のアカウントの追加可能性を判定することでアカウントの自動分類を行うことを考え、当該アカウント集合の特徴として Tweet のバースト性を抽出する手法について提案する。

2 関連研究

山本らは [1] 類似した興味を持つユーザ間ではバースト時刻が頻繁に重なることを用いて、バースト時刻によるアカウント間の類似度の算出を行っている。この手法ではバースト時間毎にトピックを LDA によって付与してアカウント間のトピックの類似度を算出している。本研究では、同じ興味を持つアカウント集合の投稿のバースト性を用いる点、バースト中の出現単語の類似度を用いる点が異なる。

3 提案手法

3.1 概要

本研究では、SNS における一定の話題を共有する初期アカウント集合が与えられたときに、他の判定対象アカウントがこの集合に属するかどうかを判定することで、初期アカウント集合を成長させることでアカウント分類を行うことを考えている。このとき、話題を共有するアカウント集合では、外部的なイベントに呼応してバースト的に投稿される現象が見られることに着目する。 個々のアカウント単位では、必ずしもバースト性を満たさないとしても、集合となったときにバースト性が見られ、一方、話題を共有していないアカウントを含めた集合ではこのバースト性は埋もれてしまうことから、これを社会的バースト性と名付ける。

本稿では、この社会的バースト性を抽出する方法を提案し、評価を行う。なお、初期アカウント集合を得る手法については議論しない *。

手順は、注目アカウント集合 A において特徴的に現れる語に重みを与え、A におけるバースト的投稿時間帯を抽出し、この重みを考慮して A に特徴的なバーストとその語集合を抽出する。その後、特徴バーストと語集合に併合されうるかを判定対象アカウント毎に判定することとなる。

3.2 語の重み

Twitter では、注目アカウント集合 A には当該タイムライン、全体アカウント集合 O には sample タイムラインを用いる。このとき、両タイムラインの同時間帯のツイート集合における名詞毎の出現頻度をそれぞれ tf_a、tf_o とする。語の重み w は、

$$w = \begin{cases} 0 & (\text{if } \mathrm{tf_o} \geq T_h) \\ \mathrm{tf_a} & (\text{otherwise}) \end{cases} \tag{1}$$

とする。T_h は適当な閾値である。

3.3 バースト区間の抽出

単位時間毎のツイート件数を元に、ピラミッド型のデータ構造を構築することで、平均到着間隔が短くなることを検知してバースト抽出を行う手法 [2] を用いる。抽出した T 個のバーストはそれぞれバースト区間 i の時間情報 $t_i(start, end)$ で識別される。

3.4 バースト区間のアカウント集合への関係度算出

各バースト区間 i のツイートでの各名詞 j の出現回数 b_{ij} と重み w_j の積の総和をバーストの得点 $\mathrm{score}(t_i)$ とする。抽出したバースト区間のうち得点の最も高いバーストと S_{burst} そのバーストに得点を与えた単語集合 S_{word}=$\{s_1, s_2, \cdots, s_n\}$ として特徴バーストを抽出する。

$$\mathrm{score}(t_i) = \sum_{j=1}^{n} b_{ij} w_j \tag{2}$$

$$S_{\mathrm{burst}} = \underset{1 \leq i \leq T}{\mathrm{argmax}}\, \mathrm{score}(t_i) \tag{3}$$

4 実験

提案手法により目的とするバーストが抽出できるかどうかを検証する。

4.1 データセット

アカウント集合として、2014 年のプロ野球日本シリーズ開幕前に「日本シリーズ」を含むツイートをしていたアカウントを人手による分類で分けた「阪神タイガースファン」の集合と「福岡ソフトバンクホーク

A social burst extraction for automatic clustering of SNS accounts.

†Satoshi Nakamura ‡Naoto Takasaki †Aki Kobayashi

†Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

‡Electrical Engineering and Electronics, Kogakuin University Graduate School

*単なる検索結果、あるいは人手による小さな集合を想定している。

スファン」の集合、「ジャイアンツ」というワードがプロフィール又はアカウント名に含まれる「読売ジャイアンツファン」の集合に対して実験を行った。「阪神タイガースファン」と「福岡ソフトバンクホークスファン」のツイートの収集は Streaming API を用い、「読売ジャイアンツファン」のツイートの収集は Rest API を用いた。それぞれのアカウント集合に対するツイート数やアカウント数などは表 1 のようである。語の重みをつける際の閾値 T_h は tf_o の上位 1 割にあたる値とした。

4.2 結果

アカウント集合それぞれの、$\mathrm{score}(t_i)$ が上位 3 位までのバースト時間は表 2 のようになった。ホークスファンのアカウント集合ではバースト回数が 1 回だけだったため結果は 1 つだけである。また、上位のバースト時間に与えた得点が高い単語をそれぞれ 3 つずつを表 3 に示す。

5 考察

それぞれのバースト時間中のツイートではそのチームの呼び方や関係のある地名が高得点を与えることが多かったので語の重みを使うのは有効であるといえる。しかし、タイガースファンのリストに登録してあるアカウントのツイートが短時間で何回もリツイートされてた結果としてバーストとして抽出しており、語の重みにおいて集合と関係ないと考えられるが、サンプルでも出現していない単語があることによって高得点を与えてしまった。リツイートを除外するとバースト回数が極端に少なくなりリプライによるバーストでスクリーンネーム (英数字で表現するアカウント名) が高得点を与えてしまっていたためリツイートやリプライの扱いについては検討する必要がある。

ホークスファンのリストではホークスが日本一になった瞬間だけがバーストとして抽出された。ジャイアンツファンのリストでは得点が最も高いバースト時間ではツイート内容はジャイアンツに関係のあるものではあったが、ニュース記事のまとめ系アカウントという単独アカウントの短時間での連続的投稿によって抽出してしまった。このことより、バーストの抽出と語の重み付けによってアカウント集合に関係のあるツイートを含む特徴的なバーストを抽出できた。

6 おわりに

アカウント集合の社会的バースト性の抽出手法について提案、検証した。リツイートや同一アカウントの短時間での多くの投稿に左右されない操作や、パラメータの検討は今後の課題である。

表 1: データセット

集合	アカウント数	収集期間	ツイート数	sample ツイート数	tf_o 閾値
タイガースファン	72	2014-10-24 22:33:22〜2015-01-08 06:43:22	80439	649824	1.37×10^{-7}
ホークスファン	50	2014-10-24 22:34:56〜2015-01-08 06:20:17	44083	649824	1.37×10^{-7}
ジャイアンツファン	100	2009-10-29 11:47:12〜2014-12-19 22:45:04	188644	1481505	6.18×10^{-8}

表 2: バースト結果

集合	順位	Score	バースト開始時間	件数	時間 (秒)
タイガースファン	1	4.784062334	2014/10/29 22:36	67	72
	2	1.0305985479	2014/10/29 22:27	71	58
	3	0.745388702	2014/10/25 16:47	40	87
ホークスファン	1	0.0106243781	2014/10/30 22:04	8	15
ジャイアンツファン	1	0.4340742957	2014/11/3 5:27	9	42
	2	0.4226758338	2014/11/23 4:07	11	183
	3	0.397137234	2014/11/26 8:04	6	2

表 3: 高得点語

集合	順位	得点 1 位	得点 2 位	得点 3 位
タイガースファン	1	嘘	fla	flap
タイガースファン	2	fla	flap	Yusa
タイガースファン	3	fla	flap	生存
ホークスファン	1	日本一	シリーズ	福岡
ジャイアンツファン	1	巨人	ジャイアンツ	読売
ジャイアンツファン	2	巨人	ジャイアンツ	野球
ジャイアンツファン	3	巨人	ジャイアンツ	野球

7 謝辞

本研究の一部は科学技術研究費 (基盤 (A))No.26242013 による。

参考文献

[1] 山本 修平, 若林 啓, 佐藤 哲司, “バースト時刻に基づくフォロー先ユーザ推定手法”, WebDB Forum 2014, A-5, 2014-11

[2] 蝦名 亮平, 中村 健二, 小柳 滋, “リアルタイムバースト検出手法の提案 ”, 日本データベース学会, 日本データベース学会論文誌. 9 巻 2 号, 1-6, 2010-11