4I1-5

ソーシャルメディアにおけるユーザーコミュニティの情報を用いたバースト予測に関する研究

Predicting the burst using an user community in social media

石塚 淳[*1] 榊 剛史[*1*2] 丸井 淳己[*1] 森 純一郎[*1] 坂田 一郎[*1]
Jun Ishitsuka Takeshi Sakaki Junki Marui Junichiro Mori Ichiro Sakata

[*1] 東京大学
The University of Tokyo

[*2] 株式会社ホットリンク
Hottolink, Inc.

Recently, we can easily translate information using a social media such as Twitter and Facebook. Because Twitter came to be used by many people, flaming came to be also paid attention. In this research, we define the burst as a phenomenon that one tweet is retweeted by many people. The result is that accuracy with the user community features is about 5% higher than that with only tweet and user features.

1. はじめに

近年, Twitter や Facebook などのソーシャルメディアと呼ばれるサービスが普及し, 一般のユーザーが容易に情報を発信できるようになった. ソーシャルメディアの中でも Twitter はリアルタイム性が高く, リツイートなどの機能によって情報伝播が起こりやすいという特徴を持っている. Twitter の普及と共に炎上というウェブ上の現象にも注目が集まっている. Twitter が普及する以前の炎上は, 芸能人のブログに対して, 閲覧者の批判的なコメントが集　　中的に集まるというような事態を指すことが多かったが, Twitter というソーシャルメディアの普及によって, 炎上は芸能人などの一部の人間だけではなく, 我々の身にも起こりうる身近なものとなっている. Twitter 上で炎上が起こっている最中では, ツイート数やリツイート数が急増することが多い. 本研究では個々のツイートに対して, そのツイートが数多くリツイートされることをバーストと定義した.

2. 関連研究

本研究では抽出したコミュニティの情報を特徴量として加えることによって, バースト予測モデルを構築する手法を提案している. 鳥海らの研究では実際の炎上事例を対象にユーザーをクラスタリングしてコミュニティを抽出し, 分析を行っている[鳥海 14]. しかし榊らの研究では, 発生した炎上事例に関わったユーザーからコミュニティを抽出しているので, 事前にコミュニティを抽出している本研究とは異なり, コミュニティ情報を予測に活かすことができない. またツイートのバースト予測モデルの構築では, Yang ら[Yang 10]や Petrovic ら[Petrovic 10]による既存研究での予測で使われているユーザー属性とツイート属性を用いて, 機械学習による分類を行う. しかしながら, 特徴量にコミュニティ属性を用いるといった研究はいまだ行われておらず, 本研究は前述した RT 予測や情報伝播予測の発展と位置付けられる.

3. 提案手法

図 1 に提案手法における全体のフレームワークを示す. 提案手法の全体の流れは, Twitter の相互メンションデータからユーザーのコミュニティを抽出し, ツイート, ユーザー, コミュニティの情報から特徴量を生成し, それらのデータに機械学習の分類モデルを適用し, ツイートのバーストを予測する二値分類器を作成するという流れになっている. コミュニティの抽出にはネットワーククラスタリングの最も代表的な手法である Louvain 法を用い, Twitter のプロフィール文からコミュニティごとに tf-idf 値が高い単語を抽出し, そのコミュニティの特徴語とした. 特徴語は LSA によって次元圧縮を行い, バースト予測モデルの特徴量に加える. バースト予測モデルには SVM を用いた.

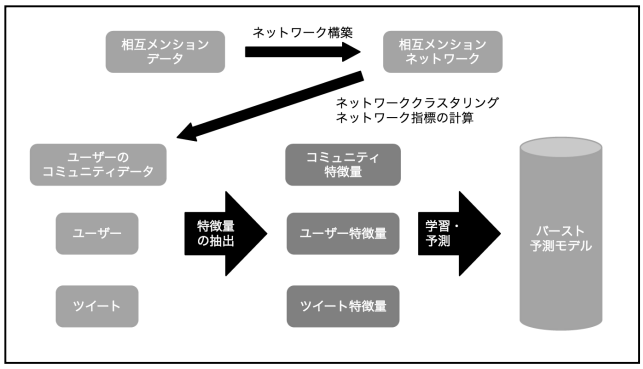

図 1. 提案手法における全体のフレームワーク

4. 実験

4.1 手順

実験では, まずモデルの考慮する特徴量と特徴語の圧縮次元数を決定するため, 3 つの事例のデータ(生活保護, 献血, 美味しんぼ)を用いて実験を行う. その後, 個別の事例に対してバースト予測モデルを構築し, 特徴量の組合せによる精度を比較する. 実験に用いた特徴量を表 1 に示す. TweetA は明確な数値で表現されるツイート属性, TweetB は内容によるツイート属性, User はユーザー属性, CommunityA はユーザーのコミュニティにおける影響度による属性, CommunityB はユーザーの所属するコミュニティによる属性である. ポジティブ率はツイートに含まれるマッチングした感情語がポジティブである確率, 主観率はツイートに含まれるマッチングした感情語が主観的である確率と定義する. マッチングの辞書には東北大学の乾・岡崎研究室の公開している日本語極性辞書 4)を用いた. また実験に用いた特徴量の組合せを表 2 に示す. ただし, 組合せ 4 は圧縮次元数を 10, 30, 50, 100, 200, 圧縮なしの 6 通り実験を行う.

連絡先:石塚 淳, 東京大学, Ishitsuka@ipr-ctr.t.u-tokyo.ac.jp

4.2 データ

実験に用いた炎上事例のツイートデータは表 3 の検索キーワードと検索期間で収集したものである. またコミュニティ抽出に用いた相互メンションデータは 2012 年 1 月 1 日から 2012 年 3 月 31 日までの 3 ヶ月間における日本国内のツイートデータから取得した相互メンションデータであり, 総リンク数は 36,743,689 本, 総ノード数は 5,980,977 となっている.

表 1. 実験に用いた特徴量

特徴量名	使用した特徴量
TweetA	ハッシュタグ数, メンション数, URL 数, 文字数, リプライか
TweetB	ポジティブ率, 主観率
User	フォロー数, フォロワー数, お気に入り数, ツイート数
Community A	クラスター係数, 次数中心性, 近接中心性, 媒介中心性, 固有ベクトル中心性, PageRank, HubScore, AuthorityScore
Community B	コミュニティのメンバー数, 特徴語

表 2. 実験に用いた特徴量の組合せ

組合せ	使用した特徴量
1	TweetA + User
2	TweetA + User + TweetB
3	TweetA + User + CommunityA
4	TweetA + User + CommunityA + CommunityB

表 3. ツイートの収集条件

事例	検索キーワード	検索期間
1	次長課長,河本,不正受給,生ポ,生活保護,ナマポ,生ぽ,なまぽ	2012/04/01 - 2012/08/31
2	庭山,献血,汚染地域	2012/04/01 - 2012/07/31
3	美味しんぼ,スピリッツ,おいしんぼ,原因不明の鼻血,風評被害,鼻血描写	2014/04/27 - 2014/04/30
4	人工知能	2013/12/16 - 2014/04/07
5	ALS,アイスバケツ,アイスバケット,"Ice bucket",氷水	2014/08/01 - 2014/09/02
6	STAP,理研,リケジョ,万能細胞,小保方,オボカタ,おぼかた,笹井,若山,バカンティ,野依	2014/01/29 - 2014/06/30

4.3 バーストの定義

本研究では, リツイート回数が 5 回以上のものをバーストしたツイート, リツイート回数が 1 回から 2 回のものをバーストしなかったツイートと定義した. バーストの定義を 5 回以上と設定した理由はリツイート 5 回以上のツイートはリツイートされたツイートの中でリツイート回数が上位 10%となっているからである. また, 本来リツイートされようのないツイートが含まれるのを防ぐため, リツイート回数が 0 回のツイートは除いた.

5. 結果と考察

まず 3 つの事例のツイートデータを利用して構築したバースト予測モデルの実験結果を表 4 に示す. 評価手法は 10-fold Cross-Validation を用いて, Accuracy を評価した. TweetB の特徴量を組合せたときのみ精度が下がったが, CommunityA や CommunityB を組合せると精度は上昇した. また次元圧縮は 10 次元が最も精度が高く, 圧縮をしないときよりも精度が上昇した.

表 4. 3 つの事例における実験結果

組合せ	Accuracy
1	59.81833
2	57.93211
3	64.53398
4(圧縮次元数 10)	65.46544
4(圧縮なし)	64.64981

次に個別の事例それぞれにおいて, バースト予測モデルを構築し, 精度の評価を行った. コミュニティの特徴量を用いると, 全ての事例で精度は上昇したが, その上昇量は事例によって異なった. コミュニティの抽出に用いたネットワークデータと同時期に炎上が起きた生活保護, 献血の事例では上昇量が大きい結果となった. コミュニティ情報は絶えず変化しているので, リアルタイムなコミュニティ情報を得ることは難しく, その点が手法の限界として挙げられる.

6. 結論

本研究では, Twitter のコミュニティを抽出し, ユーザーのコミュニティ情報を用いることでバースト予測モデルを構築した. ツイートの内容による特徴量は予測に対して有効でなかったが, コミュニティによる属性は有効であることが示された. 本研究で得られたバーストやリツイートの特徴が今後の炎上事例の鎮静や組織体系の整備に繋がれば幸いである.

参考文献

[鳥海 14] 鳥海不二夫, 榊剛史, 岡崎直観:「人工知能」の表紙に関するツイートの分析・続報, 第 4 回 Web インテリジェンスとインタラクション研究会, 2014

[Yang 10] J.Yang and S.Counts: Predicting the Speed, Scale, and Range of Information Diffusion in Twitter, In Fourth International AAAI Conference on Weblogs and Social Media, 2010

[Petrovic 10] S. Petrovic, M. Osborne, and V. Lavrenko: RT to Win! Predicting Message Propagation in Twitter, In Fifth International AAAI Conference on Weblogs and Social Media, 2010