2E1-03

テレビ番組視聴時における twitter 投稿のバースト検知と情報配信の可能性

Extract Viewer Response to the TV Program by Using Burst and Topic Detection Methods

前川浩基[*1] 中原孝信[*2] 羽室行信[*3]
Hiroki MAEGAWA Takanobu NAKAHARA Yukinobu HAMURO

[*1]（株）Magne-Max Capital Management
Magne-Max Capital Management

[*2]関西大学データマイニング応用研究センター
Data Mining Applied Recerch Center, Kansai University

[*3]関西学院大学経営戦略研究科
Institute of Business and Accounting, Kwansei Gakuin University

In this paper, we propose a new method to detect important topics from tweets on certain TV program. This method is consisted of two major parts. First, we generate small size clusters, called “micro-clusters”, by enumerating dense sub-graphs from words-relation network, which is composed of vertexes as words and edges having tight relation between two words. Second, we construct a classification model to distinguish the important clusters from the others using explanatory variables such as magnitude of burst ratio, statistics of time distribution of tweets belonging to the cluster. Finally we succeeded to build a meaningful model to retrieve the important topics.

1. はじめに

近年，インターネットを利用したコミュニケーションツールとして，マイクロブログの利用者が増加している．マイクロブログは Twitter，Jaiku，mixi ボイスなどに代表されるように，ブログとチャットの性質を併せ持ったサービスである．これらは，手軽に文章を投稿できることから，投稿までに要する時間は短く，リアルタイム性を持ったコミュニケーションツールとして利用されている．そのような特性から，マイクロブログを利用して情報を収集するユーザは多く，電車の遅延情報や震災時の安否確認などにも利用されている．また，マイクロブログの流行によって，既存のメディアからは得ることが困難であった膨大なユーザの率直な意見や，位置情報をリアルタイムに入手することが可能になった．このような情報を利用することで，特定のユーザに特定のクーポンや広告を配信したり，番組放送中に視聴者の投稿をリアルタイムで表示したり，既存のサービスやメディアとの融合が活発に行われている．

マイクロブログの普及に伴い，ソーシャルビューイング (以下 SV) と呼ばれる，テレビ番組を視聴しながらマイクロブログへ番組の感想や意見を投稿する視聴スタイルが盛んになってきている．Twitter ユーザの 54%は SV を経験しており，他人のツイートをきっかけに番組を視聴したことのあるユーザは 30.5%という調査報告 [SNS] がある．テレビを見ながら家族やお茶の間で話題を共有するというスタイルから，不特定多数の人と SNS を通じて，話題の共有や一体感を得たいという視聴スタイルへの変化が生じていると考えられる．本研究では，SV に着目し，特定の番組を視聴しながら投稿している Twitter の内容を解析することで，急激に投稿数が増加したような内容を抽出するだけではなく，番組の感想や番組に関する内容など，興味深いトピックを抽出する方法を提案する．本実験では『宇宙兄弟』を分析対象の番組として利用する．

連絡先: 前川浩基，(株)Magne-Max Capital Management，大阪市西区江戸堀 1-20-22-301，06-7176-1992，maegawa@magne-max.com

2. 関連研究

SV に関する研究は，テレビ番組を視聴しながら行われた投稿を発見することに主眼が置かれている．SV をリアルタイムに検出するために，全番組の放送タイトルや字幕スーパーを利用して特徴語を作成し，ツイート内容から作成した特徴語と一致するものを発見する。そして投稿したユーザを追跡することで，SV に関する投稿を特定している [Yamamoto 2013]．本研究では，SV としての投稿をリアルタイムに検出する必要はなく，番組の放送時間に番組に関連する内容を投稿したものを対象としている．

ニュース記事や Twitter に投稿された内容からトピックを抽出する研究は，投稿数 (文章数) の急激な増加をバーストとして検知し，トピックモデルを利用することで，バースト時に出現する単語や文章を概念化する．そして特定の話題を抽出している [Nakazawa 2011],[Takahashi 2011]．これらはいずれも Kleinberg のバースト検知 [Kleinberg 2002] を利用した方法で，ドキュメント出現数の急激な増加に着目しバーストを検知している．一方で，ドキュメントの急増を見つけるのではなく，時間区間内で出現した単語の生成確率分布からバーストしている単語を検知し，分布が類似した単語をグループ化することで，バーストイベントを抽出する方法も提案されている [Fung 2005]．これは文章ではなく，単語に着目したボトムアップ的な方法である．

これらの研究は，バーストを検知してそのトピックを発見することを目的にしているが，バースト時に出現したトピックが必ずしも興味深い話題であるとは限らない．本研究は，バースト時のトピックだけを検出するのではなく，番組の感想や番組に関する内容など，興味深いトピックの抽出を目的とする．提案方法は，Twitter に含まれる単語の共起関係を利用して，

マイクロクラスタと呼ぶ密に繋がった少数の単語からなるクラスタを生成し，それらのクラスタにラベルをつける．そして，そのクラスタを意味づける特徴量を，バースト検知や投稿間隔など様々な値を用いてモデル化する．この方法を利用することで，バースト時のトピックも含めて興味深いトピックが抽出できることを示す．

3. 手法

TV 番組についてのツイートは多様で，その出現分布も様々である．特定のトピックに一時的に多くの人が反応することもあれば，その影に隠れて比較的長時間に渡って同じトピックについて投稿されることもある．それ故に，ツイートの書き込み数が多い時間帯をバースト検知し，その時間帯のツイートを要約するだけでは見逃すトピックも多いであろう．

さらに，トピックを正しく検知できたとしても，そのトピックがユーザにとって興味あるものとは限らない．そのため，そのトピックが興味の対象であるかどうかを確認するための作業が重荷になるケースが実際には多く，理想的には，ユーザが予め定義した興味に従った内容のみを抽出できることが望ましい．以上の点を考慮し，本研究では，ユーザにとって興味深いトピックを検知するシステムの構築を目的としている．

提案手法は大きく二つのパートから構成される．まず，ツイートにおける単語の共起情報に基づいて，互いに関連の強い単語を小さなクラスタ (「マイクロクラスタ」と呼称する) として抽出する (詳細は次節)．次に，得られたマイクロクラスタに対して人間が興味深さをラベル付けすることでデータセットを作成し，クラスタのもつ各種特徴量によって興味深さを目的変数とした分類モデルを構築する．実際の運用においては，構築されたモデルを利用することで，ユーザの判断なしに，興味深いトピックを抽出することが期待できる．

興味深さの定義としては，1) 番組のストーリーに関するトピック，2) 番組についての感想/意見，の 2 つについて取り上げることにした．これらの定義については，得られたマイクロクラスタを構成するツイートを実際に著者らが読むことでラベル付けを行った．例えば，1) の定義におけるツイートは，「宇宙兄弟に千葉大出てきた」,「奥さんとの出会いはオリジナルだー。若い 2 人いいなー。奥さんいい人だー。」といったストーリーに関係のあるトピックである．また，2) に関するツイートは,「宇宙兄弟ほんと面白いね !! ここまで来ても飽きが来ない !! 」,「ケンジの今までのストーリー泣ける。」といった番組についての感想/意見である．

一方で，クラスタの特徴量としては，容易に取得/計算できる事が必要で，1) ツイートされた時刻の分布に関するもの，2) クラスタを構成する単語やツイートについての統計量，そして，3) バーストとの関係性，について表 1 に示される 6 つの特徴量を用意した．

以下では，マイクロクラスタの導出方法，および説明変数で用いているバースト検知手法について論述する．

3.1 マイクロクラスタの取得

取得したツイートから関連の強い単語をクラスタリングするために，単語を節点に，関係の強い単語に枝を張ったネットワークを構成し，そこから密な部分グラフを抽出することで，意味の近い単語のクラスタを抽出する．ただし，単語の品詞としては，動詞，形容詞，名詞，副詞，感動詞を対象とし，2 ツイート以上で出現するような単語に限定した．

表 1: 興味深さを分類する説明変数一覧

分類	説明変数	内容
ツイート内容	1) 単語数	クラスタを構成する単語数
	2) ツイート件数	クラスタを構成する単語を 2 語以上含むツイート件数 (「クラスタ」ツイートと呼ぶ)
時刻分布	3) 標準偏差	クラスタツイートの発生時刻についての各種統計量
	4) 歪度	
	5) 尖度	
バースト	6) バースト度	全ツイートについてバーストと判定した時間に含まれるクラスタツイートの割合

また，関係性の強さは PMI(pointwise mutual information) によって定義した．単語 u の生起確率を $\mathrm{p}(u)$，単語 v との共起確率を $\mathrm{p}(u,v)$ で表すと，u と v の PMI は式 (1) で定義される．

$$\mathrm{pmi}(u,v) = \log_2 \frac{\mathrm{p}(u,v)}{\mathrm{p}(u)\mathrm{p}(v)} \tag{1}$$

PMI の値が 0 より大きければ，二つの評価表現は共起しやすく，0 より小さければ共起しにくいと解釈できる．そしてユーザが指定した閾値 μ について，$\mathrm{pmi}(u,v) > \mu$ を満たすような二つの単語 u, v に枝を張る．

μ を小さな値にすると密なネットワークとなり，逆に大きな値にすると疎なネットワークが構成されることになる．本研究では $\mu = 0.5$ として実験を行った．

更に，直接の隣接関係だけでなく，間接的な隣接関係も考慮に入れることでネットワークからノイズ的な枝を除去することができ，結果として，より小さく密な部分グラフを多く含むネットワークに変換することができる．それらの密な部分グラフの抽出にはクリーク列挙を用いた．$G = (V, E)$ を節点集合 V と枝集合 E を持つ無向グラフとすると，節点集合 V の任意の節点に枝があるような G の誘導部分グラフをクリークと呼ぶ．また，あるクリークが他のクリークの真部分集合でなければ，それは極大クリークと呼ぶ．単語ネットワークから極大クリークを列挙することで，お互いに関係の強い単語集合を抽出することが可能となる．

3.2 バースト検知手法とバースト度

バースト検知には，Kleinberg により提案された手法を用いる [Kleinberg 2002]．この手法は，メッセージの到着間隔についての確率分布の変化を検出することによりバースト状態を検知する手法である．このモデルは HMM(Hidden Malkov Model) をベースにしている．HMM は系列データのモデル化手法の一つで，確率的状態遷移モデルとデータ生成モデルから構成され，観測される系列データは，隠れ状態におけるデータ生成モデルに従うと考える．

時刻 t において観測されたデータ x_t は，隠れ状態 $z_t \in \{1, 2, \cdots, K\}$ に定義された確率分布 $p(\mathbf{x}_t|z_t; \boldsymbol{\phi})$ に従って生成されるようにモデル化される．ここで，$\boldsymbol{\phi}$ は生成モデルのパラメータベクトルで，t に依存せず一定であると仮定する．

また，隠れ状態 z_t は直前の状態 z_{t-1} にのみ依存して遷移し，その確率分布は $p(z_n|z_{n-1}; \mathbf{A})$ で示される．ここで $\mathbf{A} =$

$\{a_{i,j}|i,j=1,2,\cdots,K\}$ は，状態 i から状態 j への遷移確率表で，t に依存せず一定であると仮定する．ただし，$\sum_j a_{i,j}=1.0$ で，また初期状態 z_1 は確率ベクトル $\boldsymbol{\pi}$ に従うものとする．

以上より，観測データ系列 $X=x_1,x_2,\cdots,x_T$，および状態系列 $Z=z_1,z_2,\cdots,z_T$ の同時確率は式 (2) で与えられる [Bishop 2008]．

$$p(\mathbf{X},\mathbf{Z};\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})=p(z_1;\boldsymbol{\pi})\left[\prod_{i=2}^{T}p(z_i|z_{i-1};\mathbf{A})\right]\prod_{j=1}^{T}p(x_j|z_j;\boldsymbol{\phi}) \tag{2}$$

Kleinberg のバースト検知手法は，パラメータ $\boldsymbol{\pi},A,\boldsymbol{\phi}$ が与えられたなかで，データ系列 $\mathbf{X}$ を観測した時に，式 (2) で示された同時確率を最大化するような $\mathbf{Z}$ を見つける問題として捉えることができる (式 (3))．

$$\mathbf{Z}^*=\underset{\mathbf{Z}}{\operatorname{argmax}}\,p(\mathbf{X},\mathbf{Z};\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi}) \tag{3}$$

Kleinberg のバースト検知手法では，観測データ系列 $\mathbf{X}$ がメッセージの到着間隔に対応し，隠れ状態は，定常状態とバースト状態の二状態 ($K=2$) である．そしてデータ生成モデルには指数分布 $f(x;\phi)=\phi e^{-\phi x}$ を用いている．ここで，ϕ は単位時間あたりの到着数で，定常状態における ϕ_1 は，全ツイートの到着間隔平均をとし，バースト時は $\phi_2=s\phi_1$ で与えられる．ここで，$s>0$ はスケーリングパラメータで，この値を大きく設定すれば，より際立ったバーストのみを検知することになる．

次に，遷移確率表 $\mathbf{A}$ であるが，Kleinberg の手法では，遷移確率をコスト関数に変換することで独自の最適化問題を設定しているが，本研究では，HMM の定式化に従って遷移確率そのものを与えている．実験では，定常状態からバースト状態への遷移確率 $a_{1,2}=0.3$ とし，逆の遷移確率 $a_{2,1}=0.5$ とした．そして，初期状態を決定する確率ベクトル $\boldsymbol{\pi}$ は $\pi_1=1.0$,$\pi_2=0.0$ とすることで必ず定常状態から始まるように設定される．

なお，TV 番組に対する投稿は，一般的に番組の最初と最後，そして途中の TV 広告時に増加する傾向がある．このようなデータに対してバースト検知を実施すると，それら増加時のみをバーストとして検知することになる．この問題を回避するために，宇宙兄弟の平均的な投稿分布によって投稿時刻を基準化する方法 [Fujiki 2004] を用いた．

以上の手法を宇宙兄弟が放映される 30 分間につぶやかれたツイートに適用し，全ツイートに対してバーストかどうかの判定を行い，表 1 に示されるバースト度を式 (4) に従い計算した．

$$\text{バースト度}=\frac{|W_i\cap V|/|W_i|}{|V|/|W|} \tag{4}$$

ここで，W は全ツイート集合，V はバーストと判断されたツイート集合を表し，また W_i は，あるマイクロクラスタ i におけるクラスタツイート集合である．

バースト度は，クラスタツイートにおけるバースト中につぶやかれたツイートの割合を，全体のバーストの割合で基準化したもので，この値が大きくなるほど，相対的にバースト中につぶやかれた内容を多く含むことを意味する．

4. 実験

4.1 実験データ

本研究では，TV アニメーション番組『宇宙兄弟』に関するツイートを分析対象とした．具体的には,「宇宙兄弟」「#uchukyodai」などを検索キーワードとして，2012 年 10 月 26 日から 2013 年 2 月 20 日までの約 28 万ツイートを Twitter から取得し，その中から番組放送時間中（日曜日午前 7 時〜7 時 30 分）のツイートを抽出して実験に用いた．

4.2 マイクロクラスタの生成

まず，第 35 話（「だだっ広い施設のほんの一角から」; 2012 年 12 月 2 日）放送時間中の約 1400 ツイートを用いて語のクラスタリングを行ったところ，132 のクラスタが得られた．1 クラスタあたりの語数の平均は 3.47 語，最大は 24 語であった．なおソフトクラスタリングであるため，複数のクラスタに属する語も存在する．得られたクラスタの一部と，それに付したラベルを表 2 に示す．ツイート内で共起する確率の高い語がクラスタを構成しているため，各クラスタはひとつのトピックであると考えることができる．

表 2: 各クラスタに含まれる語とラベル

No.	クラスタ	L1	L2
2	{ 日本, テレビ, 録画 }	NO	NO
58	{ 味噌汁, こぼれる }	YES	NO
71	{ 千葉, 大学, 号, 棟, 館, 中庭, 出る, 教育, 普遍 }	YES	NO
94	{ 漫画, 似る, 飽きる }	NO	YES
99	{ 奥さん, 馴れる, 始め, ほほえましい, オリジナルだ, 追加, 大幅だ, 日曜, 版, 死ぬ, なんか }	YES	YES

L1 は「番組のストーリーに関係するトピックかどうか」,
L2 は「番組についての意見/感想かどうか」を表す.

4.3 分類モデルの生成

各クラスタに付与されたラベルを目的変数とした分類モデル（決定木）を，Weka を用いて生成した.「番組のストーリーに関係するトピックどうか」の決定木を図 1 に示す.

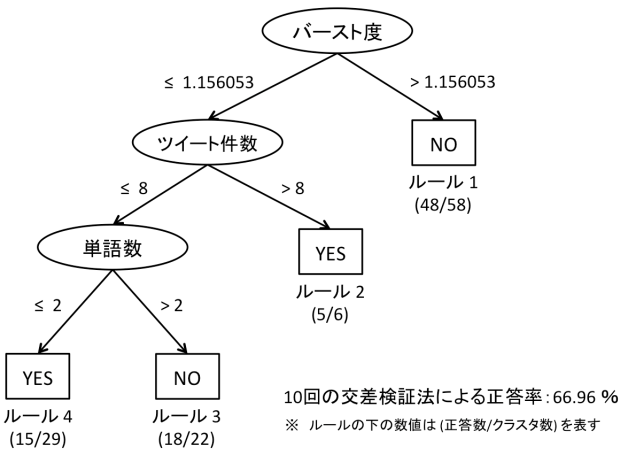

図 1: 「番組のストーリーに関係するトピック」の決定木

ルール 1 は，バースト度が高い状態を表している．バーストしているときは，番組のストーリーに関するツイートではなく，そのシーンから受けた印象や感動をツイートすることが多いため，バースト度が高いとストーリーに関係のないツイートが多く投稿される．ルール 2 は，バーストしていないがツイート件数が多い場合を示している．これは，長時間同じトピックについて語られるなど，番組のストーリーに関する投稿が多くなっていると考えられる．ルール 3 は， バースト度が低くツイート件数も少ないが，クラスタに属する単語数が多い場合を示している．たとえば { 丁寧だ, いつまでも, 止める, 続ける, いいね, 遅い } など結びつきの弱い単語が多数集まっているクラスタは解釈が難しく，ストーリーと関係しないトピックになっていると考えられる．ルール 4 は，正解率がそもそも低いため予測精度は低いが，2 語のクラスタには，{ 味噌汁, こぼれる }, { 最終, 試験 } のように特定のシーン，もしくはストーリーを特定できるものもあった．

4.4 モデルの改善に向けて

本項では，得られたモデルが分類を誤ったクラスタについて分析することで，クラスタあるいはモデルの生成における改善の方向性を検討する．

前項のモデルにおいて，誤分類をもっとも多く引き起こしていたのはルール 4 であった．この条件で YES と分類された 29 クラスタのうち，正解が YES だったのは 15 クラスタにとどまり，14 クラスタが誤分類されていた．

誤分類されたクラスタについて見てみると，{ パン, 食べる }, { 絵, 描く }, { もう, 流れる } など，共起による関係性は強いと思われるが，番組のストーリーとは無関係に出現したであろう語によるクラスタが見られた．このような理由から誤分類されたクラスタを減らすために，何らかの改善策が必要である．

なお,「番組についての感想/意見かどうか」を目的変数とした分類モデルも構築したが，そのルールはクラスタの単語数の多少だけであるという決定木が生成された（図 2）．確かに，クラスタに含まれる語数が多くなると意見・感想を含む確率は高くなるだろうが，このモデルによって視聴者の意見・感想を効果的に抽出できるとは考えにくい．語の持つ感情を数値化してモデルに与えるなどの方法を検討したい．

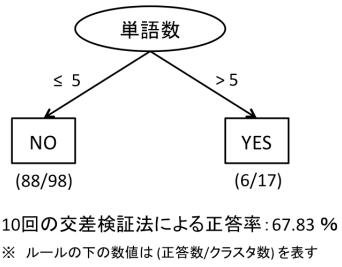

図 2: 「番組についての感想/意見かどうか」の決定木

5. おわりに

本研究は，宇宙兄弟を視聴しながらツイートした内容を対象に，マイクロクラスタリングとバースト検知を利用し，興味深いトピックを抽出するための手法を提案した．マイクロクラスタは単語間の共起関係から密に接続されたグループであり，ツイートの内容をうまく概念化しており，番組のストーリーに関係のあるクラスタと関係のないクラスタに分類することができた．特に，バースト度が低くてもツイート数が多い場合は番組のストーリーと関係のあるツイートが投稿されているというルールが発見できた．これはバーストだけに着目した場合には捉えることができないルールであり，提案した手法の有効性が確認できた．今後は，別の回の放送や，別の番組を対象に同様の実験を行い，より一般化されたルールの抽出と手法を確立していきたい．

謝辞

本研究で利用したマイクロクラスタリングは，国立情報学研究所の宇野毅明准教授が実装されたツールを利用させいただいた．本研究の一部は，ERATO 湊離散構造処理系プロジェクト，及び文部科学省の科研費若手研究 (B) 4730375 の研究助成を受けている．

参考文献

[Bishop 2008] C.M. ビショップ著, 元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇（編）, パターン認識と機械学習（下）：ベイズ理論による統計的予測, 13 章, pp.323–370, 2008.

[Fujiki 2004] 藤木稔明, 南野朋之, 鈴木泰裕, 奥村学, 「document stream における burst の発見」, 情報処理学会研究報告. 自然言語処理研究会報告, 一般社団法人情報処理学会, No.23, pp.85-92, 2004.

[Fung 2005] Fung, G., J. Yu, P. Yu and H. Lu, “Parameter free bursty events detection in text streams”, *Proceedings of the 31st international conference on Very large data bases*, No.12, pp.181-192, 2005.

[Kleinberg 2002] Kleinberg, J., “Bursty and hierarchical structure in streams”, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, No.11, pp.91–101, 2002.

[Nakazawa 2011] 中澤昌美, 帆足啓一郎, 小野智弘, 「Twitter を用いたテレビ番組からのイベント検出及びラベル付与手法」, 一般社団法人情報処理学会, pp.517-519, 2011.

[SNS] SNS × TV 連携の現状と展望 Twitter/Facebook, mixi/LINE の取り組み, `http://av.watch.impress.co.jp/docs/news/20121018_566709.html`

[Takahashi 2011] 高橋佑介, 横本大輔, 宇津呂武仁, 吉岡真治, 「ニュースにおけるトピックのバースト特性の分析」, 情報処理学会研究報告. 自然言語処理研究会報告, 一般社団法人情報処理学会,No.6, pp.1–6, 2011.

[Yamamoto 2013] 山本祐輔, 浅井洋樹, 上田高徳, 秋岡明香, 山名早人, 「テレビ番組に対する意見を持つ Twitter ユーザのリアルタイム検出」, DEIM Forum 2013 C1-4, 2013.