マイクロブログのバースト現象に着目したユーザのクラスタリングおよび可視化

大竹 洋平[1] 鈴木 良弥[2]
[1]山梨大学 工学部 コンピュータ・メディア工学科
[2]山梨大学大学院総合研究部

1 はじめに

近年,SNS の普及に伴い企業による顧客の情報活用が行われており, 自社サービスと顧客の SNS アカウントを連携したマーケティングは今後一層普及すると考えられる.

ユーザを分析する上でユーザの周囲で発生したイベントを正確に把握することが重要である. なぜならサッカーの試合などのイベント発生時には, 多くのユーザに何らかの変化が起こると考えられ, この変化が新規顧客の獲得や新製品の影響力調査, 顧客セグメントの細分化などに利用できると考えられるためである.

しかしイベントをマーケティングに活用する際, イベントを推測するための事前知識を継続的に獲得することが難しいことや利用者によって有益な情報が異なるなどの課題がある.

そこで本研究では, イベント獲得のため時系列における単語分布の急激な変化(バースト現象)に着目し, イベント単位で Twitter ユーザを集約及び可視化する手法を提案する. これによりユーザクラスタの規模やユーザ・イベント間の関係の把握が容易になり有益な情報の発見支援になると考えられる.

2 関連研究

Twitter ユーザの分類を目的とした研究として, 黒澤らの研究[1]がある. 黒澤らの手法では元投稿と返信の間で両ユーザの興味は同一であると仮定し, 投稿対に含まれる共通の特徴語に着目することで現実世界のコミュニティの特定を試みている. 本研究はユーザの分類および可視化という点で黒澤らの手法と類似しているが, イベントをクラスタリングの尺度としている点で異なる.

また, バースト現象に関する研究として文献[2][3]がある. 向井らの研究[2]ではバーストの発生タイミングに着目して情報推薦を試みている. 具体的には「文書数に基づくバースト度」の評価式[4]を改良したバースト評価式(1)を用いてバーストの検出を行っている. 高橋らの研究[3]では DTM(Dynamic Topic Model)を用いて推定したトピックに対して Kleinberg のバースト解析を適用することで, トピック単位のバースト検出が可能であることを示している.

本研究では高橋らの手法にならい, トピック単位でバーストの検出を行うが, 獲得したバースト語集合をイベント毎に集約する点, ユーザの分類を目的とする点で先行研究とは異なる. また本研究は向井らの考案したバースト検出の手法をトピックモデルに適用した研究であると言える.

3 提案手法

3.1 イベント集合の抽出

① 言語データ

Wikipedia 辞書エントリを追加した MeCab により形態素解析を行い名詞の抽出を行った. その際, 引用 URL とハッシュタグ, ユーザ名, RT などの文字列および 1 文字で表記される名詞や Wikipedia にエントリが存在しない名詞はノイズであると考えられるため抽出の対象から除外した.

② 潜在的ディリクレ配分法(LDA)

本研究では LDA(Latent Dirichlet Allocation)[6][7]を用いて, バースト語の集合をトピックのバーストとして検出および取得する. なお今回の実験ではハイパーパラメータ α, β およびトピック数 K をそれぞれ α = 0.05, β = 0.100, K = 30. Gibbs サンプリングの試行回数を 1000 回とした.

③ バースト検出

本研究ではバーストの検出にバースト評価式(1)を用いる. 向井らはバースト判定値の算出につぶやき数を用いているが, 本研究ではトピックの生起確率を用いてバースト判定値を算出する. 評価式中の X は 3, Y については解析期間全体における解析日の前日までの全ての日数に設定し, 判定値 B が閾値 1.0 を超えた時バースト状態だとみなす.

$$B = \frac{N}{\sqrt{\bar{A}}} \cdot \frac{N-A}{N+A} \quad \cdots(1)$$

N :その区間におけるトピックの生起確率
A :直前X区間のトピックの生起確率の平均
$\bar{A}$:直前Y区間のトピックの生起確率の平均

図 1 バースト評価式

3.2 投稿ベクトルの重みづけ

本研究ではバースト検出したトピックの生起確率上位語をバースト語とする．またバースト語がユーザの分類に寄与するようにユーザ毎に重み付け（出現回数×30）を行った．なお，今回の実験では上位 20 語のうち人手で選出した 12 語を重みづけの対象とした．

3.3 pLSA による次元圧縮

計算時間を短縮するため Hoffman による pLSA（Probabilistic Latent Semantic Analysis）[5][8]を用いて 15 次元に次元圧縮を行った．なお温度パラメータ β は 0.75 を採用した．

3.4 自己組織化マップによる視覚化

投稿ベクトルの視覚化に Kohonen による自己組織化マップ（Self-Organizing Map）を使用する．自己組織化マップはニューラルネットワークの一種であり，多次元ベクトルデータを学習し，類似した性質を持つデータが近接するように 2 次元平面上へ写像する．ユーザの位置関係によりユーザやイベント間の関係，クラスタの規模などが直観的に把握できるため知見発見の支援になると考えられる．なお，今回の実験で設定したマップサイズ，学習回数，初期学習率係数 α，初期近傍半径 r を表 1 に示す．

表 1 SOM のパラメータ

マップサイズ	12ノード×8ノード	
1st	学習回数	1000
	学習率係数 α	0.05
	近傍半径 r	10
2nd	学習回数	10000
	学習率係数 α	0.02
	近傍半径 r	3

4 実験条件

今回の実験では頻繁にバーストが発生すると考えられるスポーツを対象に興味があると思われるユーザ 38 人を人手で収集し，その投稿を実験データとした．なおバーストの解析日と解析期間を表 2 に示す．

表 2 バースト解析日と解析期間

バースト解析期間	解析日	解析間隔
2014/11/07 ～ 2014/11/21	2014/11/19	1日

5 実験結果と考察

5.1 トピックのバースト検出

解析期間における投稿に対して LDA を用いてトピック推定を行った．表 3 に人手で付与したトピックのラベルと各トピックの特徴語，表 4 に解析期間中に観測した各日のイベントを示す．表 3,4 から各日の特徴語をトピックに集約できていると言える．

次に 17～19 日に高い生起確率を示した 3 つのトピックと定常的に出現する語からなるトピック 1 つに対してトピック分布の時間推移を可視化した結果とバースト検出を行った結果をそれぞれ図 2, 3 に示す．図 2, 3 からイベントの発生日に，その日の特徴語が集約されたトピックがバーストしていることがわかる．また「www」のように定常的に出現する語からなるトピックが期間全体を通してバーストしていない様子がわかる．以上のことから，各日におけるイベント集合をトピックのバーストとして検出できたと考えられる．またバースト検出したトピックの単語分布を単語のバースト度の尺度とすることでバースト語の抽出が可能であることが示せた．一方，抽出したバースト語の中には「アア」のようなノイズも見受けられた.この結果は同一人物による連続投稿や連続表記,あるいはデータ数の不足による偶発的なバーストによるものだと考えられる．

表 3 トピック別の特徴語

トピックのラベル	各トピックの特徴語
11月15日	ノーヒットノーラン， ノーノー， ng， シャーレ， 讃岐， 台湾， 失点， 継投， mlb， samurai
11月16日	バンビ， エリザベス女王杯， ob， 台湾， 松井， 錦織圭， 錦織， ラキシス， 引退， 巨人阪神
11月17日	流星群， 内定， 日テレ， 真琴， オス， 鳩山， 解説， cd， 取消， ホステス
11月18日	日本， サッカー， オーストラリア， キリンチャレンジカップ， vs， ゴール， 本田， 日本代表， 岡崎， 代表
11月19日	西島秀俊， アア， 結婚， am， 復活， 紹介， 流行語大賞， ノミネート， 俳優， 没収試合
11月20日	ベストナイン， 今宮， ヤクルト， 阿部， bs， 銀次， オリックス， 進撃の巨人， 決定， 配役
定常的な発言	www， ww， 選手， 今日， rt， 野球， ちゃん， 試合， 明日， ファン

表 4 トピックに出現した日付別の話題

日付	トピックに出現した日付別の出来事
11月15日	日米野球：侍ジャパンノーヒットノーラン達成
11月16日	エリザベス女王杯，男子テニスATPツアー・ファイナル準決勝，巨人vs阪神OB戦
11月17日	おうし座&しし座ダブル流星群，日テレ女子アナ内定取消
11月18日	キリンチャレンジカップ：日本代表vsオーストラリア代表
11月19日	西島秀俊の結婚発表，流行語大賞ノミネート語発表，アギーレJ初白星没収試合
11月20日	プロ野球ベストナイン発表，「進撃の巨人」の配役決定

5.2 自己組織化マップによる集約結果

実験結果を図 4 に示す. 図は隣接ノード間の距離を濃淡で表現した図であり, 白で表現されたノードほど距離が近く, 同じような性質を持つノードで構成された領域である. なお今回の実験ではクラスタの特徴を把握しやすくするためにメディアンフィルターを施して平滑化を行い, ユーザが存在するノードに対しては表 5 に示すユーザのラベルと人数を付与している. 図 4 を見ると, 右上に西島英俊の結婚発表に関するクラスタが形成され, 中央部には流行語大賞ノミネート語やアギーレジャパン没収試合などのイベントに関するクラスタが形成されていることが分かる. このことから特定のユーザに対してイベントという単位でユーザの集約ができたと考えられる. 一方, 誤分類されるユーザも見受けられた. この結果は複数のイベントに関する投稿を行ったことによる特徴の分散, あるいはバースト語の重み付けや pLSA の次元数が不適切だった可能性がある. また, 投稿数が少なく特徴に乏しいユーザが不適切な次元に圧縮される問題や無属性のユーザから構成されるクラスタが散在するといった問題は今後の課題である.

表 5 ユーザのラベルとイベント

ラベル	イベント
A	西島秀俊の結婚発表
B	流行語大賞ノミネート語発表
C	アギーレJ初白星没収試合
D	紹介（偶発的なバースト語）
E	無属性（イベントなし）

6 参考文献

[1]黒澤義明,竹澤寿幸, マイクロブログサービスの返信行動に着目した投稿及びユーザの分類, 言語処理学会 第 17 回年次大会 pp.460-463

[2]向井友宏,黒澤義明,目良和也,竹澤寿幸,マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案, 言語処理学会 第 17 回年次大会 pp.452-455

[3]高橋佑介,横本大輔,宇津呂武仁,吉岡真治,河田容英,神門典子,福原知宏,中川裕志,清田陽司,時系列トピックモデルにおけるバーストの同定,DEIM2012

[4]八村太輔,湯本高行,赤星祐平,小山聡,田中克己,Web 検索結果のクラスタリングと観点抽出に基づく閲覧インタフェース,DEWS2005, 4B-i11(2005)

[5]Hofmann, T. (1999). ”Probabilistic Latent Semantic Indexing.”in Proceedings of the 22nd Annual International ACMSIGIR Conference on Research and Development in Information Retrieval”, pp.50-57.

[6]D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[7]GibbsLDA++, http://gibbslda.sourceforge.net/

[8]工藤拓. “PLSI”, http://www.chasen.org/~taku/software/plsi/

[9]som_pak, http://www.cis.hut.fi/research/som_pak

[10]形態素解析エンジン MeCab http://mecab.sourceforge.net/

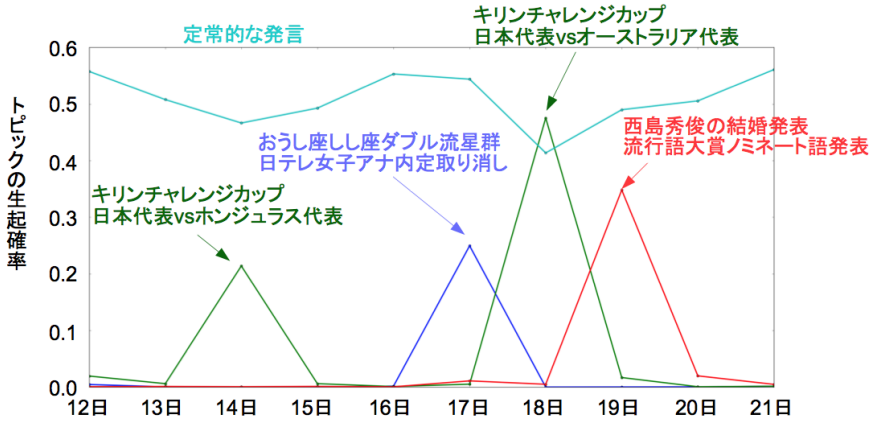

図 2　トピック分布の時間推移

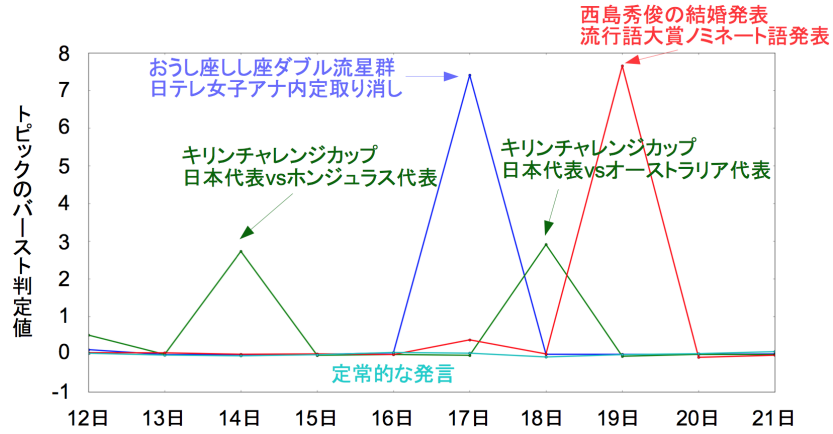

図 3　トピック別のバースト判定値と時間推移

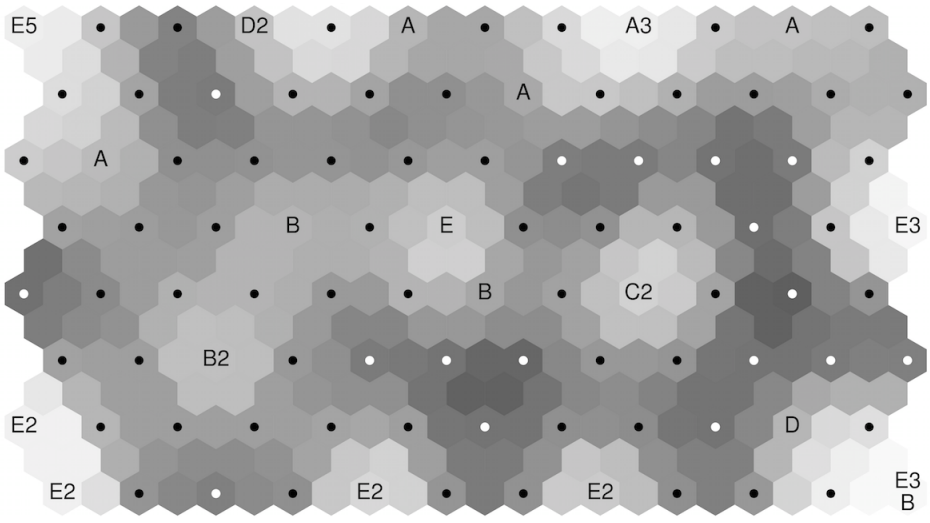

図 4　自己組織化マップによるユーザの集約結果