# 時系列ニュースにおけるトピックのバーストの同定*

高橋 佑介†　横本 大輔†　宇津呂 武仁‡　吉岡 真治§
筑波大学大学院 システム情報工学研究科†
筑波大学 システム情報系‡　北海道大学大学院 情報科学研究科§

## 1 はじめに

現代の情報社会においては，多種多様な情報が氾濫し，いわゆる情報爆発の問題が深刻であり，氾濫する情報の集約や，俯瞰を行うための技術の確立が強く望まれている．中でも，情報爆発が最も顕著に現れているのはウェブであり，ウェブ上の情報爆発の問題に取り組んだ研究が盛んに行われている．例えば，バースト解析の技術においては，ストリームデータの時間軸方向の密度から世の中の異変や特異な出来事を捉えることができる．また，別のアプローチとして，トピックモデルのように文書集合における主要なトピックを推定することのできる技術も存在する．

バースト解析は，一般には，電子メールやウェブ上のニュース記事のようなストリームデータに対して適用される．そこでは，ある時からある話題に関する記述が急激に増加するような現象が起こることがあり，こういった現象を，ある話題に関するバーストと呼ぶ．代表的なアルゴリズムであるKleinbergのバースト解析 [5] では，時系列に沿った各キーワードのバースト度の変化や，バーストしているか否かの判定，バースト度によるキーワードのランク付けをすることができる．

一方，トピックモデルにおいては，文書が生成される背景には，潜在的にいくつかのトピックがあることを想定し，文書の生成尤度を高めるようにモデルのパラメータを訓練する．トピックモデルの一種であるDTM (dynamic topic model) [3] においては，時系列情報を持つ文書集合を情報源として，時系列にそって，各単位時間ごとに，文書ごとのトピックの分布と，トピックごとの語の分布を求めることができる．

以上をふまえて，本論文では，キーワードではなくトピックを対象としてバースト解析を行うことを目的とする．具体的には，DTMによって解析期間におけるトピックの分布を推定し，提案手法に基づいて各トピックの関連文書数を定義する．これにより，トピックに対してKleinbergのバースト解析手法が適用できるようになる．実際に，ウェブ上の時系列ニュースを対象にして本手法を適用することにより，特定の期間において集中して記事が観測されるチリ地震やバンクーバー五輪に関するバーストを，トピックの単位で検出することができるようになった．

## 2 Kleinbergのバースト解析アルゴリズム

本研究では，Kleinbergの考案したバースト解析アルゴリズム [5] を用いた．このアルゴリズムを用いることで，文書ストリーム中のあるキーワードのバースト期間と非バースト期間とを自動で切り分け，各キーワードに対してバースト度を付与することが可能になる．

なお，Kleinbergは，2種類のバースト解析手法を提案しているが，本研究ではenumeratingバーストのアルゴリズムを利用する．enumeratingバーストのアルゴリズムは，離散時間で送られる文書の集合に対して適用される．本稿では，各日ごとのニュース記事集合を一つの文書集合の単位とし，以下では単に，記事集合と呼ぶ．

最も簡単なモデルでは2状態オートマトン $\mathcal{A}^2$ を定義し，2つの状態を非バースト状態 $q_0$，バースト状態 $q_1$ とおく．入力に対して状態が遷移することにより，2つの状態を切り分ける．本論文では，特定のキーワードを含む記事を「関連記事」，そうでない記事を「非関連記事」と設定し，一日の記事集合における関連記事の割合に基づいて，バースト状態であるか否かの判定を行う．

解析期間において，$m$ 個の記事集合 $B_1, \ldots, B_m$ が離散時間で送られてくる状況を考える．$t$ 番目の記事

*Identifying Bursts of Topics in Time Series News
†Yusuke Takahashi, Daisuke Yokomoto, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba
§Masaharu Yoshioka, Graduate School of Information Scienceand Technology, Hokkaido University

集合を $B_t$ とし，その記事集合に含まれる記事の数を $d_t$ とおく．文書集合には関連記事と非関連記事が含まれ，$B_t$ に含まれる関連記事の数を $r_t$ とおく．解析期間における全ての記事の数 $D$ は $D = \sum_{t=1}^{m} d_t$，解析期間における全ての関連記事の数 $R$ を $R = \sum_{t=1}^{m} r_t$ と表すことができる．

次に，オートマトンの 2 状態にそれぞれ期待値を割り当てる．初期状態である非バースト状態 $q_0$ には，解析期間全体から算出した期待値 $p_0 = R/D$ を割り当てる．バースト状態 $q_1$ には，$p_0$ にパラメータ $s$ をかけた値である $p_1 = p_0 s$ を割り当てる．ただし，$s > 1$ であり，$p_1 \leq 1$ となるような $s$ でなくてはならない．$s$ の値が小さいほど，記事集合中の関連記事の割合が低くてもバーストと見なされやすくなる．

解析は，$m$ 個の記事集合が与えられたときの，状態の系列を通るためのコスト計算によって行う．考えられる状態の系列のうち，最も系列のコストが小さいものが解となり，その系列の状態に応じて，バースト期間と非バースト期間を決定する．

状態遷移は $d_t$ と $r_t$ が入力となって決まる．状態の系列は $\mathbf{q} = (q_{i_1}, \ldots, q_{i_m})$ と表され，$q_{i_m}$ は，$m$ 番目の記事集合によって決定された状態 $q_i$ $(i = 0, 1)$ である．記事集合中の関連記事が二項分布 $B(d_t, p_i)$ にしたがって現れるという考えに基づき，状態 $q_i$ にいることに対してコストを与える関数 $\sigma(i, r_i, d_t)$ を以下のように定義する．

$$\sigma(i, r_t, d_t) \quad = \quad -\ln\Big[ \binom{d_t}{r_t} p_i^{r_t} (1 - p_i)^{d_t - r_t} \Big]$$

ただし，閾値付近の入力が続くなどして頻繁に状態遷移が起こると，途切れ途切れにバースト状態と非バースト状態が切り替わり不自然である．そこで，現在の状態 $q_i$ から次の状態 $q_j$ へ，状態遷移を妨げるための関数 $\tau(i, j)$ を定義する．

$$\tau(i, j) \quad = \quad \begin{cases} (j - i)\gamma & (j > i) \\ 0 & (j \leq i) \end{cases}$$

$\tau$ は，パラメータ $\gamma$ によって調節されるが，特に理由がない場合は $\gamma = 1$ とする．

以上に述べた，ある状態 $q$ にいることに対してコストを与える関数 $\sigma$ と，状態遷移にペナルティを課す関数 $\tau$ を使って，状態の系列 $\mathbf{q}$ を通るためのコスト関数を定義する．

$$c(\,\mathbf{q}\,|\,r_t, d_t\,) \quad = \quad \Big( \sum_{t=0}^{m-1} \tau(i_t, i_{t+1}) \Big) + \Big( \sum_{t=1}^{m} \sigma(i_t, r_t, d_t) \Big)$$

オートマトン $\mathcal{A}^2$ は二つのパラメータ $s$，$\gamma$ によって決まることから，$\mathcal{A}^2_{s,\gamma}$ と表記される．本実験では，$s = 2$，$\gamma = 1$ として $\mathcal{A}^2_{2,1}$ のオートマトンを用いている．

なお，本論文では 1 日ごとにバースト度を算出しているため，$t_k = t_l\,(= t)$ である．したがって，その際のバースト度は次のように表すことにする．

$$bw(t, w) \quad = \quad bw(t, t, w)$$

## 3 トピックモデル

本研究では，トピックモデルとして DTM (dynamic topic model) [3] を用いる．DTM は，語 $w$ の列によって表現される時間情報を含んだ文書の集合と，トピック数 $K$ を入力とし，各単位時間について，各トピック $z_n$ $(n = 1, \ldots, K)$ における語 $w$ の確率分布 $p(w|z_n)$ $(w \in V)$，及び，各文書 $b$ におけるトピック $z_n$ の確率分布 $p(z_n|b)$ $(n = 1, \ldots, K)$ を推定する．ここで，$V$ は文書中に出現する語の集合である．

DTM は，潜在的ディリクレ配分法 (LDA, Latent Dirichlet Allocation) [4] とは異なり，文書集合中の時系列情報を考慮しているため，日付等の単位時間を超えて同一トピックを追跡可能である．

本論文では，$p(w|z_n)$ $(w \in V)$，及び，$p(z_n|b)$ $(n = 1, \ldots, K)$ の推定においては，Blei らによって公開されたツール[1]を用いた．ハイパーパラメータ $\alpha$ と，トピック数 $K$ は，それぞれ $\alpha = 0.01$，$K = 20$ とした．

## 4 トピックモデルのバースト解析

Kleinberg のバースト解析は，各日における文書数 $d_t$ と，その日の関連文書数 $r_t$ を入力として，解析期間におけるバースト状態と非バースト状態を切り分けて出力する手法である．したがって，Kleinberg の手法を用いてトピックのバーストを測るためには，各日における各トピックの関連文書数 $r_t$ が得られれば良い．そこで，本手法ではトピック $z_n$ の関連文書数 $r_t$ を以下のように定義することで，トピックのバースト解析を行う．

$$r_t = \sum_{b} p(z_n|b)$$

これより，解析期間における全ての関連記事数 $R = \sum_{t=1}^{m} r_t$ が求まり，それを解析期間における全ての記事

[1] http://www.cs.princeton.edu/~blei/topicmodeling.html

表 1: DTM によって推定されたトピック (3 月 1 日時点)

| 人手でトピックに付与したラベル | $p(w\|z)$ の高いキーワード（上位 10 キーワード） |
|---|---|
| 経済 | ドル，ユーロ，上昇，市場，動き，相場，円高，売買，発表，取引 |
| トヨタリコール事件 | トヨタ，リコール，問題，社長，公聴会，トヨタ自動車，米国，無償，中国，加速 |
| バンクーバー五輪 | 選手，女子，日本，バンクーバー，3 月 1 日，日本時間<br>バンクーバー五輪，銀メダル，大会，男子 |
| 自然現象 | 津波，チリ，メートル，被害，午後，沿岸，午前，地震，センチ，発生 |
| 日本の政治 | 首相，政府，予算，国会，民主党，政策，衆院，法案，審議，方針 |

の数 $D = \sum_{t=1}^{m} d_t$ で割ることにより，解析期間全体における期待値 $p_0 = R/D$ を算出する．

## 5 分析

対象とした解析期間は，2010 年 2 月 1 日 ～ 3 月 31 日である[2]．

はじめに，解析期間におけるニュース記事データに対して DTM を用いてトピック推定を行った．代表的なトピックについて，人手で付与したトピックのラベル，および，各トピックについて $p(w|z)$ の高いキーワードのうちの上位 10 キーワードを表 1 に示す．なお，DTM では各日ごとにトピックがもつ語の確率分布が変化するが，今回の解析期間において $p(w|z)$ の高い上位 10 キーワードは各日ごとに大きく変化しなかったため，表 1 においては，2010 年 3 月 1 日時点のキーワードを示す．

次に,「バンクーバー五輪」,「経済」の 2 つのトピックに対して，本手法を用いてバーストの同定を行った結果を図 1，および， 図 2 に示す．図 1 からは，バンクーバー五輪の開催期間中に正しくトピックのバーストが検出されている様子がわかる．一方，図 2 からは．「経済」のように毎日定常的に報道されるトピックが，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，本手法を用いることにより，時系列ニュースにおいて，あるトピックについての報道が頻繁に行われている期間を，トピックのバーストとして検出可能であることが分かる．各日におけるバースト同定結果が適切であるか否かについて人手で評価を行った結果を表 2 に示す．

[2]日経新聞 (http://www.nikkei.com/)，朝日新聞 (http://www.asahi.com/)，読売新聞 (http://www.yomiuri.co.jp/) の各新聞社のサイトから収集した 6,710 記事，10,976 記事，および，9,117 記事の合計 26,896 記事．

表 2: バースト同定結果に対する人手の評価

| 期間 | 検出したバーストの数 | 正解数 | 適合率 |
|---|---|---|---|
| 2010 年 2 月 1 日 ～3 月 31 日 | 109 | 82 | 0.75 |

## 6 関連研究

文献 [8] では，キーワードのバースト度から，各日におけるトピックのバースト度を算出する手法を提案している．また，ある日においてトピックがバースト状態あるか否かの判定は，その日のバースト度が閾値を超えたか否かによって行なっている．そのため，人手でバースト度の閾値を調整する手順を経る必要がある点が短所となる．一方，本研究は，バースト度についての閾値を人手で調整する必要がない点が有利である．

文献 [7, 6] においては，Kleinberg のバースト解析手法を用いて選定したバーストキーワードに対して，トピックへの集約を行う枠組みを提案している．しかし，これらは本研究とは異なり，DTM や LDA 等のトピックモデルを用いていない．文献 [7] ではバースト度の高い上位 20 キーワードを含む文書をクラスタリングし，その結果を基に，話題ごとのキーワードの集約を行なっている．一方，文献 [6] では，共起度によってバーストキーワードを集約したものをトピックとしており，トピックのバースト度は，集約されたキーワードの中で，そのうち最もバースト度の高いキーワードのバースト度を採用している．

本研究では，トピック同士を比較する尺度としてバースト度を用いたが，文献 [2] では，トピックモデルにおいて意味のないトピック (J/I; Junk/Insignificance Topic) の語の分布を定義し，LDA によって推定されたトピックと J/I との分布間の距離を測ることでトピック同士を比較する手法を提案している．

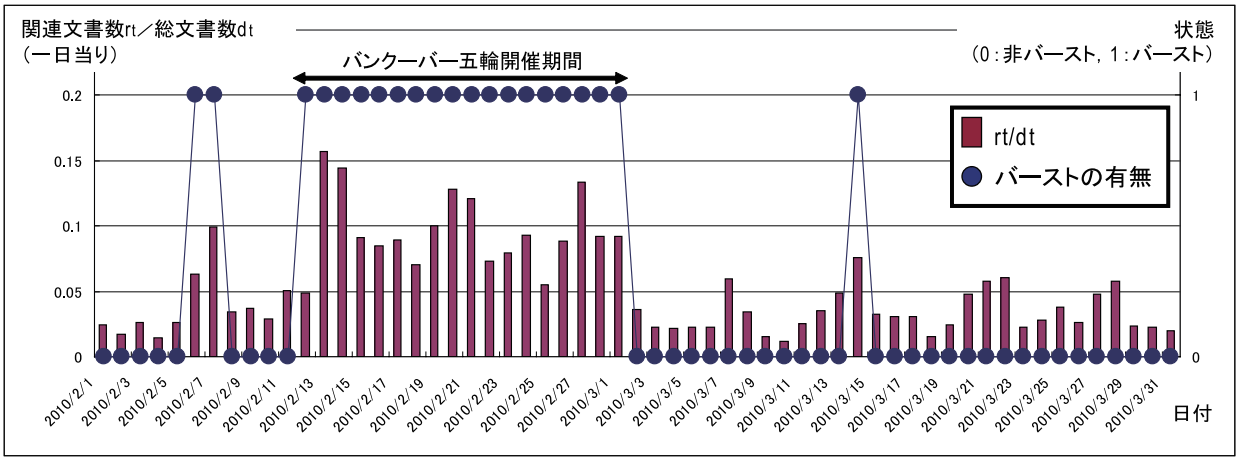


図 1: トピック「バンクーバー五輪」におけるバーストの同定結果

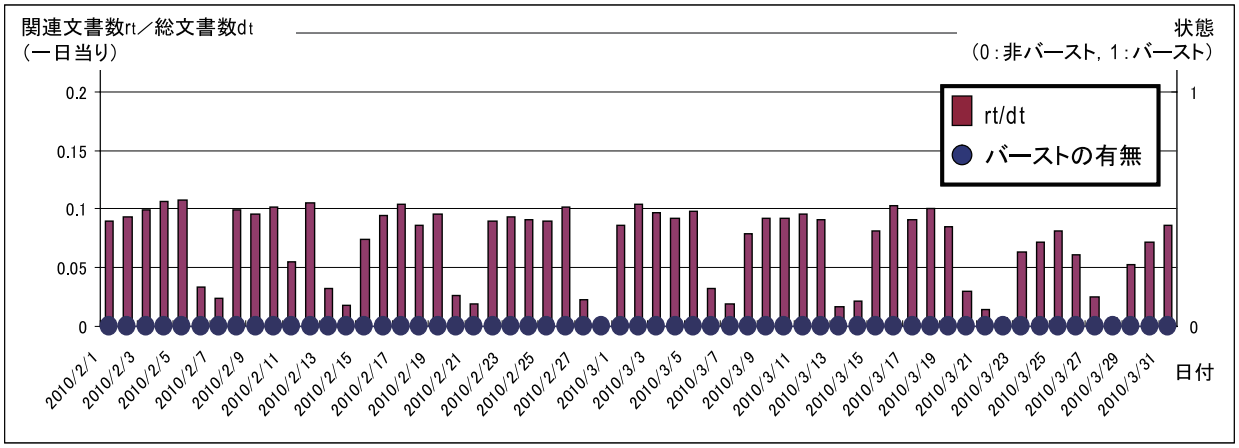


図 2: トピック「経済」におけるバーストの同定結果

## 7　おわりに

本論文では，DTM によって時系列ニュースにおけるトピックを推定し，それらのトピックの関連文書数を定義することにより，Kleinberg のバースト解析アルゴリズムを用いてトピックのバーストの同定を行う手法を提案した．これにより，キーワードに比べてより情報の単位が大きいトピックのバーストをとらえることが可能になり，時系列ニュースにおけるトピックの特徴やトピック同士の相関関係をいっそう明らかにできることを示した．今後は，トピック推定の段階でバーストの同定を行うモデルを開発する．また，Online LDA [1] などを利用し，本手法のオンライン化に取り組む．

## 参考文献

[1] L. AlSumait, D. Bardara, and C. Domeniconi. On-Line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proc. 8th ICDM*, pp. 3–12, 2008.

[2] L. AlSumait, D. Bardara, J. Gentle, and C. Domeniconi. Topic significance ranking of LDA generative models. In *Proc. ECML/PKDD*, pp. 67–82, 2009.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proc. 23rd ICML*, pp. 113–120, 2006.

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

[5] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th SIGKDD*, pp. 91–101, 2002.

[6] K. Mane and K. Borner. Mapping topics and topic bursts in PNAS. In *Proc. PNAS*, Vol. 101, Suppl 1, pp. 5287–5290, 2004.

[7] 高橋佑介, 宇津呂武仁, 吉岡真治. ニュースにおけるバーストキーワードの話題への集約. 第 3 回 DEIM フォーラム論文集, 2011.

[8] 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, Vol. 2011, No. (2011–NL–204), 2011.