DEIM Forum 2012 F5-5

時系列トピックモデルにおけるバーストの同定

高橋　佑介[†1]　横本　大輔[†1]　宇津呂武仁[†2]　吉岡　真治[†3]　河田　容英[†4]
神門　典子[†5]　福原　知宏[†6]　中川　裕志[†7]　清田　陽司[†7]

†1 筑波大学大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
†2 筑波大学システム情報系知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
†3 北海道大学大学院 情報科学研究科　〒 060-0808 北海道札幌市北区北 8 条西 5 丁目
†4 (株) ナビックス　〒 141-0031 東京都品川区西五反田 8-3-6
†6 独立行政法人 産業技術総合研究所 サービス工学研究センター　〒 135-0064 東京都江東区青梅 2-3-26
†5 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2
†7 東京大学 情報基盤センター　〒 113-0033 東京都文京区本郷 7-3-1

あらまし　本論文では，時系列ニュースを対象として，情報集約を行うための二種類の方式として，バースト解析およびトピックモデルの 2 つの手法の考え方を組み合わせることにより，トピックのバーストを検出する方式を提案する．時系列ニュースにおけるバーストとは，世の中における特異な出来事に対応して，ある時期からその出来事に関連するニュース記事が急激に増加する現象を指す．バーストを検出するための代表的な手法として，Kleinberg のバースト解析が挙げられる．この手法においては，一般的に，バーストの検出はキーワード単位で行われる．一方，文書集合におけるトピックの分布を推定するものとして LDA (latent Dirichlet allocation) や DTM (dynamic topic model) に代表されるトピックモデルがある．トピックモデルを適用することにより，ニュース記事集合全体の情報を，いくつかのトピックに集約することができる．以上の既存技術をふまえて，本論文では，DTM を用いて推定したトピックに対して Kleinberg の手法を適用することで，トピック単位のバーストが検出可能であることを示す．
キーワード　時系列ニュース, トピック，バースト, 集約

Identifying Bursts in Time Series Topic Model

Yusuke TAKAHASHI[†1], Daisuke YOKOMOTO[†1], Takehito UTSURO[†2], Masaharu YOSHIOKA[†3], Yasuhide KAWADA[†4], Noriko KANDO[†5], Tomohiro FUKUHARA[†6], Hiroshi NAKAGAWA[†7], and Yoji KIYOTA[†7]

†1 Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba, 305-8573, Japan
†2 Faculty of Engineering, Information and Systems, University of Tsukuba,　Tsukuba, 305-8573, Japan
†3 Grad. Sch. of Information Science and Technology, Hokkaido University,　Sapporo, 060-0808, Japan
†4 Navix Co., Ltd.　Tokyo 141-0031, Japan
†6 Center for Service Research, National Institute of Advanced Industrial Science and Technology, Tokyo, 135-0064, Japan
†5 National Institute of Informatics, Tokyo 101-8430, Japan
†7 Information Technology Center, University of Tokyo, Tokyo 113-0033, Japan

Key words　time series news, topic, burst, aggregation

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

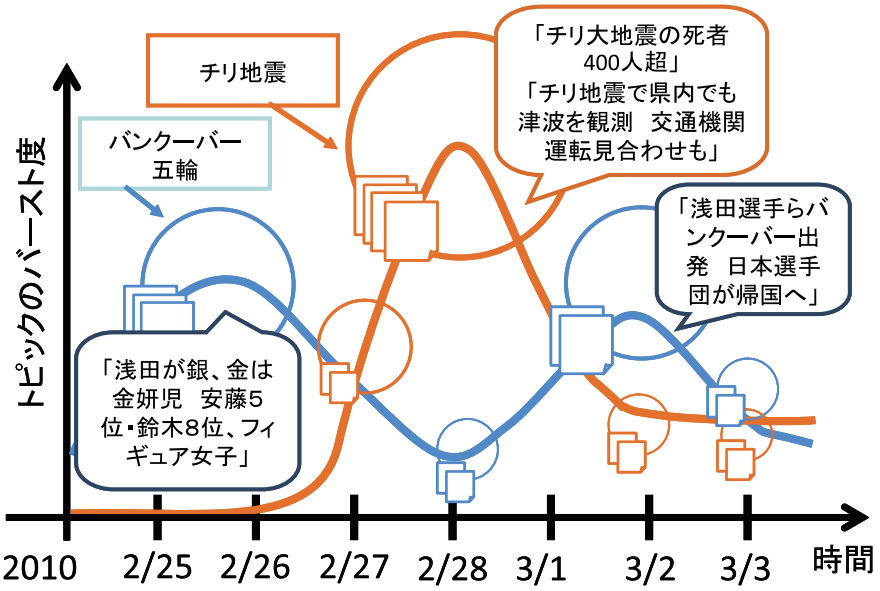

図 1　時系列ニュースにおけるトピックのバースト

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，図 1 に示すように，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．

2.1　enumerating バースト

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．目的とする記事(注1)を「関連記事」，そうでない記事を「非関連記事」として扱い，バーストか否かは，記事集合中の関連記事の割合によって決まる．

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．

(注1)：例えば，特定のキーワードを含む記事．

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，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
小沢一郎違法献金疑惑	民主党，北教組，選挙，自民党，参院選，幹事長，議員，政治，資金，事件
海外の政治	中国，大統領，米国，日本，政府，北朝鮮，関係，交渉，禁煙，国際
企業	社長，企業，販売，会社，開発，生産，日本，発表，大手，工場
企業の業績	10，カ月，発表，連続，価格，01，26，12 月，20，00
交通	運転，事故，キロ，合わせ，午後，午前，自転車，東京，メートル，時間
裁判	被告，判決，裁判員，裁判，事件，被害者，裁判長，男性，懲役，検察
スポーツ，製品情報	ネット，サイト，インターネット，発売，携帯電話，パソコン，サービス，機能，発表，情報
普天間問題	知事，問題，政府，米軍，沖縄，首相，受け，沖縄県，外相，普天間
芸能	監督，映画，作品，放送，舞台，東京，人気，ドラマ，午後，主演
刑事事件	容疑者，容疑，逮捕，事件，女性，捜査，県警，男性，死亡，発表
地域	販売，合わせ，イベント，店舗，商品，人気，発売，作品，デザイン，東京
学校，コラム	自分，子ども，生徒，学校，参加，家族，ながら，児童，受け，しょう
社会	対象，受け，調査，採用，制度，年度，学生，以上，企業，期間
医療	病院，患者，受け，医師，医療，研究，教授，検査，発表，手術
地方の行政	年度，市長，地域，計画，施設，予算，議会，県内，自治体，整備

ただし，$s > 1$ であり，$p_1 \leqq 1$ 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．

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

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．

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

τ は，パラメータ γ 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 σ と，状態遷移にペナルティを課す関数 τ 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．

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

オートマトン $\mathcal{A}^2$ は二つのパラメータ s，γ 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．

2.2　キーワードのバースト度

Kleinberg のバーストアルゴリズムでは，ある期間における各キーワードのバーストの強さを表す尺度としてバースト度を用いる．

期間 $t_k, \ldots, t_l$ におけるキーワード w のバースト度 $bw(t_k, t_l, w)$ は以下の式で定義される．

$$bw(t_k, t_l, w) \;=\; \sum_{t=t_k}^{t_l} \big(\sigma(0, r_t, d_t) - \sigma(1, r_t, d_t)\big)$$

なお，今回は 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．

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

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(注2)を用いた．ハイパーパラメータ α 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方式

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．

また，これにより期間 $t_k, \ldots, t_l$ におけるトピック z_n のバースト度 $bz(t_k, t_l, z_n)$ を以下の式により算出することができる．

$$bz(t_k, t_l, z_n) = \sum_{t=t_k}^{t_l} \big(\sigma(0, r_t, d_t) - \sigma(1, r_t, d_t)\big)$$

なお，今回は 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．

$$bz(t, z_n) = bz(t, t, z_n)$$

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(注3)．

5.1 DTM によるトピックの推定

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．表 1 に，人手で付与したトピックのラベルと，各トピックについて $p(w|z)$ の高いキーワードのうち上位 10 キーワードを示す．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．

表 2　バースト同定結果に対する主観評価

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

5.2 トピックのバーストの同定

「バンクーバー五輪」,「自然現象」,「経済」,「社会」の 4 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 2 ～ 図 5 に示す．図 2 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．また，図 3 からは，地震などが発生した際に,「自然現象」のトピックがバーストしていることがわかる．一方，図 4, および，図 5 からは.「経済」,「社会」のよう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．

以上のことは，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を示している．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主観評価を行った結果を表 2 に示す．

5.3 トピックのバースト度

2010 年 2 月 1 日 ～ 3 月 4 日の期間において,「トヨタリコール事件」,「小沢一郎違法献金疑惑」,「バンクーバー五輪」,「自然現象」の 4 つのトピックのバースト度 $bz(t, z_n)$ をプロットしたものを図 6 に示す．

図より，2 月 5 日付近では,「小沢一郎違法献金疑惑」,「トヨタリコール事件」のバースト度が他のトピックよりも高くなっていることがわかる．しかし，バンクーバー五輪が開催された 2 月 13 日付近からは,「バンクーバー五輪」がもっともバースト度の高いトピックなっており，さらにその後の 3 月 1 日付近では，2 月 27 日にチリで発生した大地震の影響により,「バンクーバー五輪」を抑えて「自然現象」のトピックが上位になっている様子がわかる．このことからわかるように，トピックのバースト度の尺度を用いることで，バーストするトピック間の優劣を定量化することができ，さらにその優劣が時系列方向に変動する様子を観測することができる．

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．具体的には，DTM を用いて各日のトピックごとの各キーワードの条件

(注2)：http://www.cs.princeton.edu/~blei/topicmodeling.html
(注3)：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．

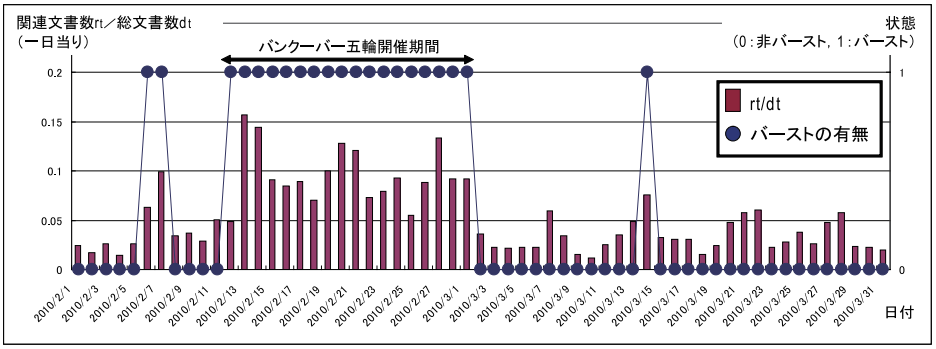

図 2　トピック「バンクーバー五輪」におけるバーストの同定結果

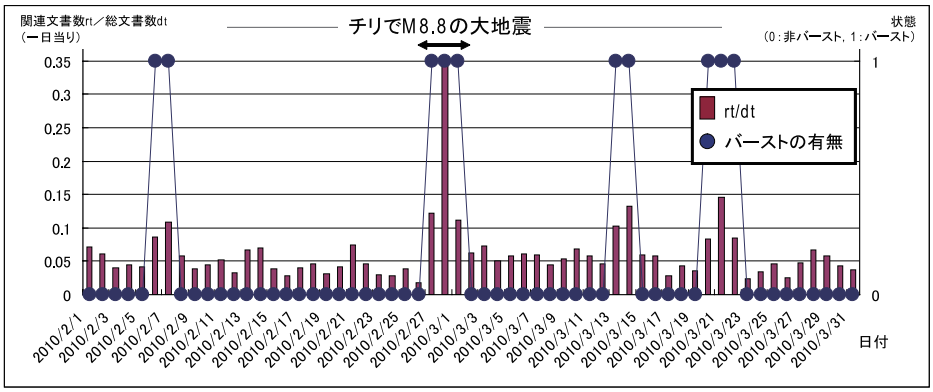

図 3　トピック「自然現象」におけるバーストの同定結果

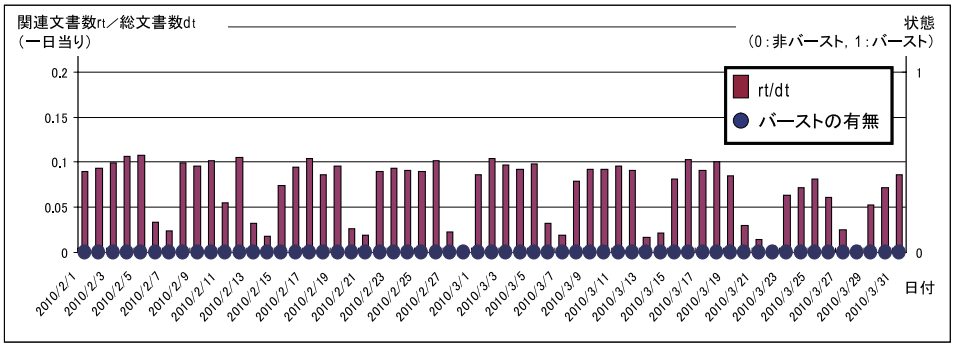

図 4　トピック「経済」におけるバーストの同定結果

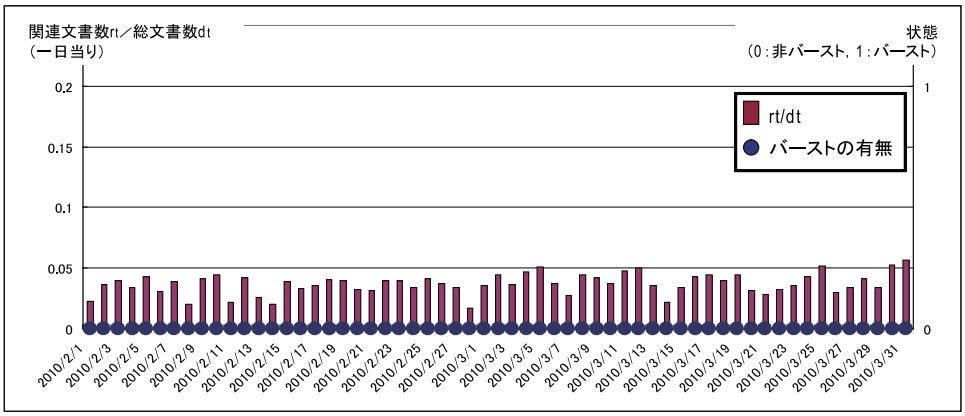

図 5　トピック「社会」におけるバーストの同定結果

付き確率と，その日におけるキーワードのバースト度との積の和を求める．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

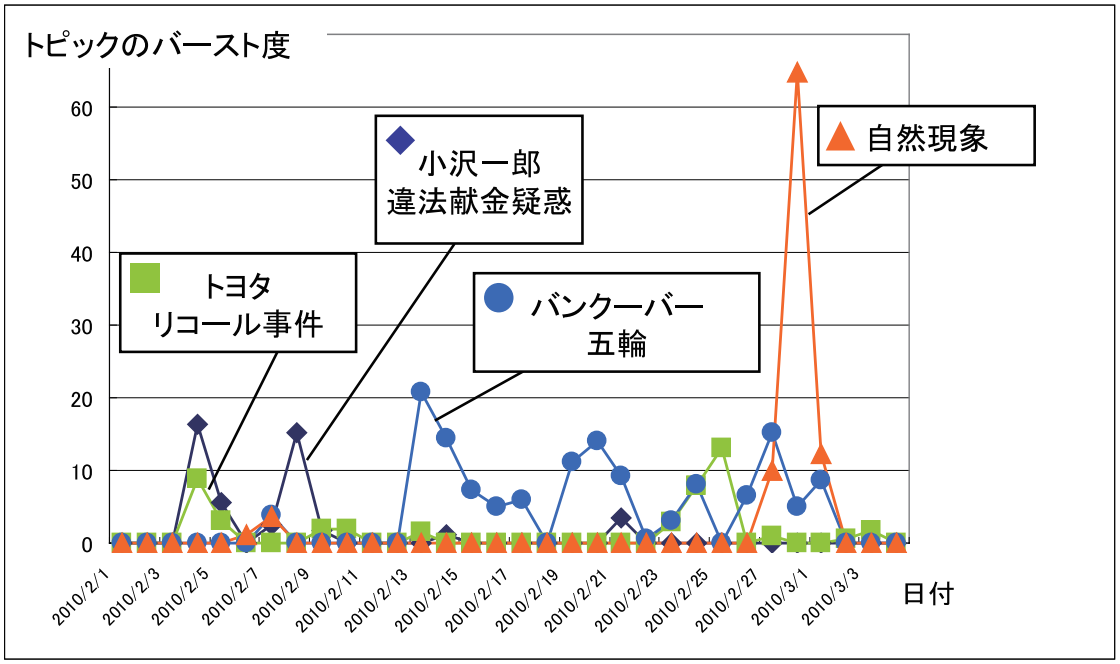

図 6　トピックのバースト度の時系列推移

調整する手順を経る必要がある点が短所となる．一方，本研究では，トピックの関連文書数を定義することにより，Kleinberg の提案するキーワードのバースト解析を行う枠組みをそのまま用いてトピックのバーストを同定する．そのため，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 [6], [7] 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，トピックのバースト度やトピック間の関係性をグラフで視覚的に表示する手法を提案している．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．

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．提案手法の詳細な評価項目として，バースト期間の長短とバースト同定精度との間の相関について分析するとともに，適合率だけでなく再現率の評価を行う枠組みを実現する．また，トピックモデルを適用する際のトピック数として，多様な粒度でのトピック推定を行った後，それぞれのトピック数のもとでのバースト同定を行うことにより，階層的なトピックにおけるバースト同定性能の評価を行う．

本論文では，まずトピックモデルによってトピックを推定し，それを対象として Kleinberg のバースト解析を行っているが，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，On-line LDA [1] などを利用し，本手法のオンライン化についても取り組む．

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.