DEIM Forum 2013 C2-4

# ニュース・ツイッター間の対応を考慮したバースト・トピックの同定

高橋　佑介†　　胡　　碩†　　宇津呂武仁††　吉岡　真治†††　神門　典子††††

† 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
†† 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
††† 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
†††† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2

**あらまし**　本論文では，時系列ニュース，および，ツイッターを混合した時系列の文書ストリームを対象として，教師なしの時系列トピックモデルを適用することにより，時系列のトピックの推移を推定する．次に，時系列に推移するトピックにおいて，ニュース，および，ツイッターの文書ストリームの分離を行った上で，Kleinberg のバースト解析モデルを適用する．この方式により，ニュース・ツイッターの混合文書ストリームに対して時系列トピックモデルを推定したにも関わらず，ニュース，および，ツイッターに対して，時系列トピックにおけるバーストが別々に同定できることを示す．

**キーワード**　トピック・バースト, 時系列ニュース, ツイッター, 集約, Kleinberg, トピックモデル, dynamic topic model, 複数情報源

# Identifying Bursty Topics considering Correlation of Time Series News and Twitter

Yusuke TAKAHASHI†, Shuo HU†, Takehito UTSURO††, Masaharu YOSHIOKA†††, and Noriko KANDO††††

† Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba, 305-8573, Japan
†† Faculty of Engineering, Information and Systems, University of Tsukuba,　Tsukuba, 305-8573, Japan
††† Grad. Sch. of Information Science and Technology, Hokkaido University,　Sapporo, 060-0808, Japan
†††† National Institute of Informatics, Tokyo 101-8430, Japan

**Abstract**　This paper first studies how to apply a time series topic model to the document stream of the mixture of time series news and twitter. Next, we divide news streams and twitter into distinct two series of document streams, and then we apply our model of bursty topic detection based on the Kleinberg's burst detection model. With this procedure, we show that, even though we estimate the time series topic model with the document stream of the mixture of news and twitter, we can detect bursty topics independently both in the news stream and in twitter.

**Key words**　topic burst, time series news, twitter, aggregation, Kleinberg, topic model, dynamic topic model, multiple information source,

## 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

表 1　評価用文書集合

| 文書集合 | ニュース | ツイッター | 合計 |
|---|---|---|---|
| ニュース記事 (全記事)<br>＋ ツイート (ロンドン五輪) | 全 11,202 記事(注1)<br>(2012/07/24 ～ 2012/08/13) | ロンドン五輪に<br>関連する<br>57,414 ツイート<br>(2012/07/24~2012/08/13) | 68,616 文書 |
| ニュース記事 (ロンドン五輪)<br>＋ ツイート (ロンドン五輪) | ロンドン五輪に関連する 2,308 記事 (注2)<br>(2012/07/24 ～ 2012/08/13) | | 59,722 文書 |

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) [2] 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，本論文の前段の研究 [9] 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手法を提案した．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した．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った．

この方式をふまえて，本論文では，密接に関連しあう 2 種類の情報源から推定した時系列トピックの分布を推定し，情報源ごとに独立にトピックのバーストを同定する機能を実現した．実際に，この機能を，時系列ニュース記事とツイートの 2 種類を混合した文書集合に対して適用することにより，2 つの情報源に密接に関連するトピックのバーストや，片方の情報源のみに現れるトピックのバーストを容易に観測することができるようになった．

## 2. 評価用文書集合

本研究では，時系列ニュース記事とツイートの 2 種類を混合した文書集合に対して評価を行う．表 1 に，評価用文書集合の内訳を示す．

### 2.1 ニュース記事

ニュース記事は，朝日新聞(注3)，日経新聞(注4)，読売新聞(注5)の各新聞社のサイトから収集した．表 1 の評価用文書集合の内訳に示すように，ニュース記事集合は，収集対象期間の全記事から構成される集合，および，そのうちの「ロンドン五輪」に関連する部分集合 (括弧書きで「ロンドン五輪」と書かれたもの) の二通りから構成される．ここで,「ロンドン五輪」に関連するニュース記事集合は，人手で選定したロンドン五輪に関連する 8 キーワード(注6) を 1 つ以上含むニュース記事から構成される．

(注3)：http://www.asahi.com/
(注4)：http://www.nikkei.com/
(注5)：http://www.yomiuri.co.jp/
(注6)：五輪，ロンドン，オリンピック，金メダル，銀メダル，銅メダル，選手，日本代表

### 2.2 ツイート

ツイートは，ツイッター(注7)社から提供されている Streaming API を利用して収集したツイート 9,509,774 件のうち，公式リツイート，および，本文中に URL を含まないツイート 7,752,129 件を利用した．この際，リプライ，ハッシュタグ付きツイート，他人のツイートを引用したツイートについても，全て通常のツイートと同様に扱う．そして,「ロンドン五輪」に関連するニュース記事絞り込みの場合と同様に，人手で選定したロンドン五輪に関連する 8 キーワードを 1 つ以上含むツイート 57,414 ツイートのみに限定した．

## 3. 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ることが可能になる．なお，文献 [5] では，2 種類のバースト解析手法が提案されているが，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．本論文では，各日ごとの文書集合を一つの文書集合の単位とする．

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．目的とする文書(注8)を「関連文書」，そうでない文書を「非関連文書」として扱い，バーストか否かは，文書集合中の関連文書の割合によって決まる．

解析期間において，$m$ 個の文書集合 $B_1, \ldots, B_m$ が離散時間で送られてくる状況を考える．$t$ 番目の文書集合を $B_t$ とし，その文書集合に含まれる文書の数を $d_t$ とおく．文書集合には関連文書と非関連文書が含まれ，$B_t$ に含まれる関連文書の数を $r_t$ とおく．解析期間における全ての文書の数 $D$ は $D = \sum_{t=1}^{m} d_t$，解析期間における全ての関連文書の数 $R$ 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，$p_1 \leqq 1$ となるような $s$ でなくてはなら

(注7)：https://twitter.com/
(注8)：例えば，特定のキーワードを含む文書.

ない．$s$ の値が小さいほど，文書集合中の関連文書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，$m$ 個の文書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，$m$ 番目の文書集合によって決定された状態 $q_i$ $(i = 0, 1)$ である．文書集合中の関連文書が二項分布 $B(d_t, p_i)$ にしたがって現れるという考えにもとづき，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．

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

$\tau$ は，パラメータ $\gamma$ によって調節される．

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．

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

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．

## 4. 時系列トピックモデルの適用

### 4.1 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) [2] 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 $p(w|z_n)$ $(w \in V)$，および，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) [3] 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，$p(w|z_n)$ $(w \in V)$，および，$p(z_n|b)$ $(n = 1, \ldots, K)$ の推定においては，Blei らによって公開されたツール(注9)を用いた．

### 4.2 適 用 手 順

本論文では，表 1 に示すニュース記事とツイートの混合文書集合に対して DTM を適用する．ここで，DTM によりトピックモデルを推定する際には，ニュース記事およびツイートの間の違いは考慮されないため，2 種類の文書集合に渡って共通のトピックモデルが推定される．本論文においては，この方式を用いることにより，ニュース記事とツイートの間で対応付けられたトピックモデルの推定を実現している．

なお，トピックモデルの推定において用いる語としては，日本語版 Wikipedia(注10)のエントリタイトル，および，リダイレクト名を使用した．また，ハイパーパラメータ $\alpha$ は $\alpha = 0.01$ とし，トピック数 $K$ は，文書集合「ニュース記事 (全記事)+ ツイート (ロンドン五輪)」のとき $K = 70$，文書集合「ニュース記事 (ロンドン五輪)+ ツイート (ロンドン五輪)」のとき $K = 50$ とした．

## 5. トピックに対するバーストの同定方式

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．この考え方にもとづき，文献 [9] 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，トピックに対するバーストの同定を行った．

$$r_t = \sum_b p(z_n|b)$$

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．

## 6. 混合文書集合に対して推定されたトピックに対するバーストの同定方式

本研究では，混合文書集合から得られたトピックについて，2 種類の時系列文書ごとに独立にバーストの同定を行う．そのため，トピックモデル推定時においては 2 種類の文書を区別なく扱ったが，トピックのバースト解析時においては，混合文書集合を 2 種類の時系列文書集合に分離して扱う．

具体的には，前節のトピック・バーストの同定方式をふまえて，密接に関連する 2 種類の時系列文書 $b_x$，$b_y$ から構成される混合文書集合に対してトピックモデルが推定されたとして，このトピックを対象として，2 種類の時系列文書ごとに独立に

(注1)：朝日新聞，日経新聞，読売新聞の内訳はそれぞれ，3,458 記事，4,587 記事，および，3,157 記事．

(注2)：朝日新聞，日経新聞，読売新聞の内訳はそれぞれ，970 記事，679 記事，および，659 記事．

(注9)：http://www.cs.princeton.edu/~blei/topicmodeling.html

(注10)：http://ja.wikipedia.org/

表 2　バースト同定結果に対する評価結果 (ニュース記事 (全記事) + ツイート (ロンドン五輪), ニュース・ツイッター個別, 全 70 トピック)

(a) ニュース

| フィルタリング | バースト解析のパラメータ | 検出したバーストの数 | 正解数 | 適合率 (%) | 正解数 ／「正解数の上限値」(%) |
|---|---|---|---|---|---|
| 無し | $s = 4$, $\gamma = 3$ | 56 日 / 45 バースト | 23 日 / 16 バースト | 41.1 (日単位) / 35.6 (バースト単位) | 48.9 (日単位) / 42.1 (バースト単位) |
| | $s = 3$, $\gamma = 2$ | 147 日 / 129 バースト | 47 日 / 38 バースト (「正解数の上限値」と表記) | 32.0 (日単位) / 29.5 (バースト単位) | 100 (日単位) / 100 (バースト単位) |
| 有り | $s = 4$, $\gamma = 3$ | 5 日 / 5 バースト | 4 日 / 4 バースト | **80.0** (日単位) / **80.0** (バースト単位) | 8.51 (日単位) / 14.3 (バースト単位) |
| | $s = 3$, $\gamma = 2$ | 31 日 / 23 バースト | 16 日 / 12 バースト | 51.6 (日単位) / 52.2 (バースト単位) | 34.0 (日単位) / 31.6 (バースト単位) |

(b) ツイッター

| フィルタリング | バースト解析のパラメータ | 検出したバーストの数 | 正解数 | 適合率 (%) | 正解数 ／「正解数の上限値」(%) |
|---|---|---|---|---|---|
| 無し | $s = 3$, $\gamma = 2$ | 62 日 / 39 バースト | 45 日 / 26 バースト | 72.6 (日単位) / 66.7 (バースト単位) | 81.8 (日単位) / 86.7 (バースト単位) |
| | $s = 2.5$, $\gamma = 1$ | 130 日 / 97 バースト | 55 日 / 30 バースト (「正解数の上限値」と表記) | 42.3 (日単位) / 30.9 (バースト単位) | 100 (日単位) / 100 (バースト単位) |
| 有り | $s = 3$, $\gamma = 2$ | 42 日 / 25 バースト | 38 日 / 23 バースト | **90.5** (日単位) / **92.0** (バースト単位) | 69.1 (日単位) / 76.7 (バースト単位) |
| | $s = 2.5$, $\gamma = 1$ | 61 日 / 34 バースト | 47 日 / 27 バースト | 77.0 (日単位) / 79.4 (バースト単位) | 85.5 (日単位) / 90.0 (バースト単位) |

バーストを同定する．ここで，トピック $z_n$ について，情報源 $x$ および $y$ に対応する関連文書数 $r_{t,x}$，$r_{t,y}$ を以下のように定義する．

$$r_{t,x} = \sum_{b_x} p(z_n|b_x)$$

$$r_{t,y} = \sum_{b_y} p(z_n|b_y)$$

これより，情報源 $x$ および $y$ に対して，解析期間における全ての関連記事数 $R_x = \sum_{t=1}^{m} r_{t,x}$ ，$R_y = \sum_{t=1}^{m} r_{t,y}$ がそれぞれ求まる．これらを，解析期間における全ての記事数 $D_x$，$D_y$ でそれぞれ割ることにより，解析期間全体における期待値 $p_{0,x} = R_x/D_x$，$p_{0,y} = R_y/D_y$ をそれぞれ算出する．

以上の定式化によって，2 つの情報源 $x$ および $y$ から生成された文書から構成される混合文書集合に対して推定されたトピック $z_n$ について，2 種類の時系列文書に対してそれぞれ独立にバーストの同定が行えるようになる．

なお，本論文においては，$x$，および，$y$ を，それぞれ，時系列ニュース記事，および，ツイートとして，以上の定式化を適用する．

## 7. 評　　価

### 7.1 トピックのバーストの同定結果の評価

DTM によって得られたトピックに対して，本手法を用いてトピックのバーストの同定を行い，バーストの正否の評価を行った．バースト正否評価の際の評価基準としては，以下の両方が満たされた場合に，トピックのバーストとして適切であると判定した(注11)．

(i)　そのトピックに対応するニュース記事あるいはツイートの集合には，話題のまとまりがある．

(ii)　バーストが同定された期間において，ニュース記事あるいはツイートで言及された事項が，実際にバーストであると判定できる．

この評価結果の詳細を表 2，および，表 3 に示す．この評価結果の要点を列挙すると，以下のようになる．

(a)　特に，ニュース記事においては，文書集合「ニュース記事 (ロンドン五輪)+ ツイート (ロンドン五輪), 2012 年 7~8 月」の方が，文書集合「ニュース記

(注11)：ロンドン五輪に関係のないバーストであっても，バーストが同定された期間において，当該トピックに関連のある話題のまとまりがあり，バーストしていると判断できれば正解とした．

表 3　バースト同定結果に対する評価結果 (ニュース記事 (ロンドン五輪) + ツイート (ロンドン五輪), ニュース・ツイッター個別, 全 50 トピック)

(a) ニュース

| フィルタリング | バースト解析のパラメータ | 検出したバーストの数 | 正解数 | 適合率 (%) | 正解数 ／「正解数の上限値」(%) |
|---|---|---|---|---|---|
| 無し | $s=4$, $\gamma=3$ | 67 日 / 52 バースト | 18 日 / 8 バースト | 26.9 (日単位) / 15.4 (バースト単位) | 54.5 (日単位) / 40.0 (バースト単位) |
| | $s=3$, $\gamma=2$ | 197 日 / 171 バースト | 33 日 / 20 バースト (「正解数の上限値」と表記) | 16.8 (日単位) / 11.7 (バースト単位) | 100 (日単位) / 100 (バースト単位) |
| 有り | $s=4$, $\gamma=3$ | 20 日 / 9 バースト | 18 日 / 8 バースト | **90.0** (日単位) / **88.9** (バースト単位) | 54.5 (日単位) / 40.0 (バースト単位) |
| | $s=3$, $\gamma=2$ | 63 日 / 46 バースト | 33 日 / 20 バースト (「正解数の上限値」と表記) | 52.4 (日単位) / 43.5 (バースト単位) | 100 (日単位) / 100 (バースト単位) |

(b) ツイッター

| フィルタリング | バースト解析のパラメータ | 検出したバーストの数 | 正解数 | 適合率 (%) | 正解数 ／「正解数の上限値」(%) |
|---|---|---|---|---|---|
| 無し | $s=3$, $\gamma=2$ | 52 日 / 28 バースト | 48 日 / 26 バースト | **92.3** (日単位) / **92.9** (バースト単位) | 51.6 (日単位) / 47.1 (バースト単位) |
| | $s=2$, $\gamma=1$ | 143 日 / 98 バースト | 93 日 / 53 バースト (「正解数の上限値」と表記) | 65.0 (日単位) / 54.1 (バースト単位) | 100 (日単位) / 100 (バースト単位) |

表 4　ニュース・ツイッター共通/固有のバースト同定精度（全 50 トピック中，ロンドン五輪に関係し，話題のまとまりがある 34 トピックが対象）

| | 共通のバーストの同定精度 (正解数/システムの出力数) | 各情報源固有のバーストの同定精度 (正解数/システムの出力数) |
|---|---|---|
| ニュース | 日単位: 87.5 % (14/16) | 日単位: 100 % (2/2) ，トピック単位: 100 % (1/1) |
| ツイッター | トピック単位: 87.5 % (7/8) | 日単位: 100 % (32/32) ，トピック単位: 100 % (13/13) |

事 (全記事)+ ツイート (ロンドン五輪), 2012 年 7~8 月」の場合よりも高い性能を達成している．この理由は，文書集合「ニュース記事 (ロンドン五輪)+ ツイート (ロンドン五輪), 2012 年 7~8 月」において，記事数の少ないトピックのバーストを自動的に除去できているため，誤同定となるバーストが削除済みだからである．一方，文書集合「ニュース記事 (全記事)+ ツイート (ロンドン五輪), 2012 年 7~8 月」においては，全体の記事数が多いため，記事数の少ないトピックの中に，話題としてのまとまりを持ったトピックが混在しており，これらのトピックに対する誤答が目立っているためである．

(b)　文書集合「ニュース記事 (ロンドン五輪)+ ツイート (ロンドン五輪), 2012 年 7~8 月」においては，ニュース記事において，記事数の少ないトピックが混在しており，このトピックのバーストを除去するためのフィルタが不可欠である．ニュース記事においては，フィルタによりこれを除去することにより(注12)，最も高い性能を達成している．

(c)　性能を下げる要因となるトピックの多くは，ロンドン五輪とは無関係なトピックが雑音となっている場合が多い．

### 7.2　ニュース・ツイッター間のバースト・トピックの比較

バースト・トピックについて，情報源間の関係性を評価するため，ニュースとツイッターのバーストの同定結果を比較する．バーストの同定結果については，最も高い適合率を達成した設定 (文書集合「ニュース記事 (ロンドン五輪)+ ツイート (ロンドン五輪)」の場合 (表 3) において，ニュース記事側では，記事数の少ないトピックを自動的に除去した場合 (フィルタリング有り)，かつ，バースト解析のパラメータが $s=4$，$\gamma=3$ の場合，および，ツイート側では，バースト解析のパラメータが $s=3$，$\gamma=2$ の場合) を用いる．

(注12)：具体的には，全文書中において当該トピックに対応する文書の割合が1%未満の場合に，そのトピックのバーストを除去した．

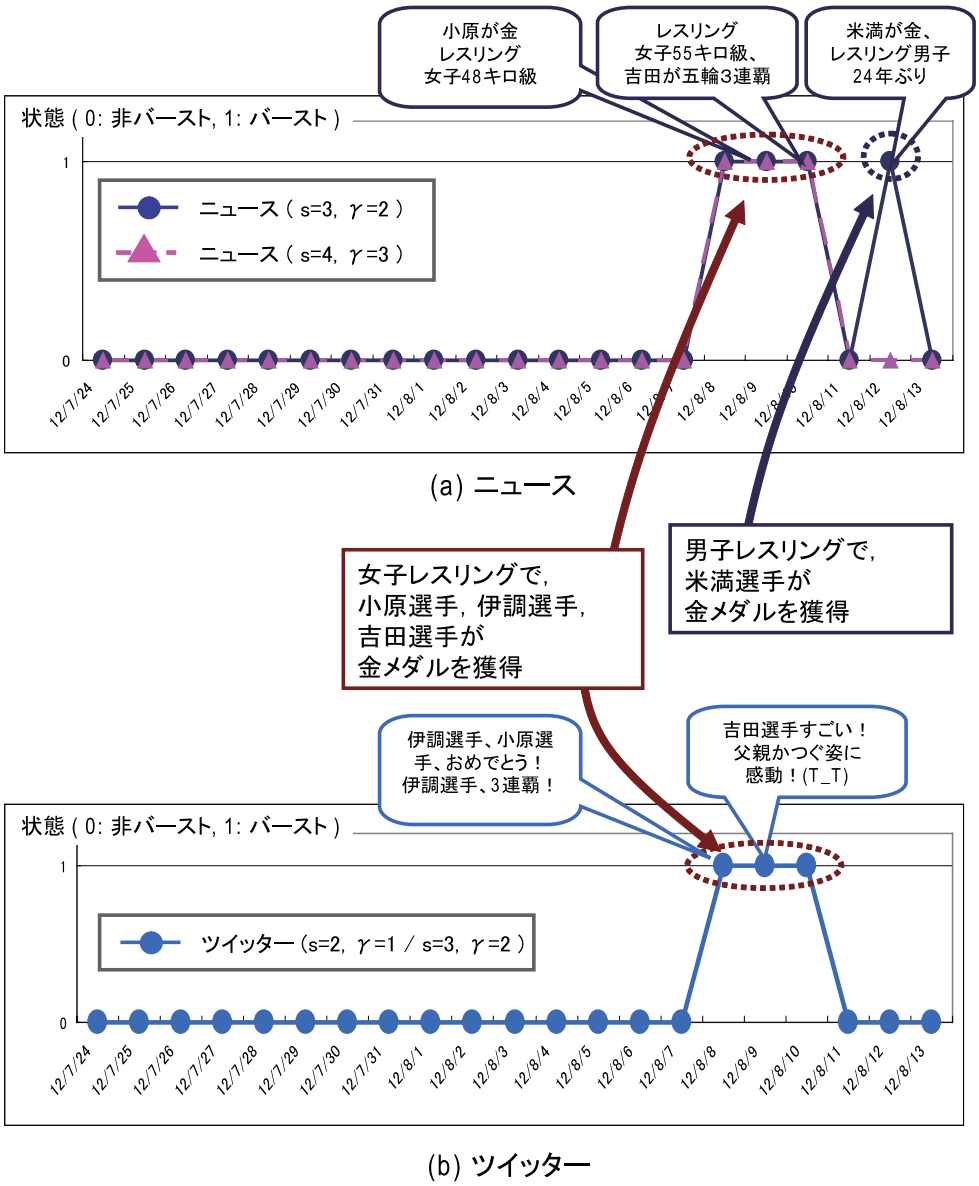



図 1　トピック「レスリング」におけるバーストの同定結果

まず，全 50 トピックのうち，ロンドン五輪に関係があり，話題としてまとまりがあると判断した 34 トピックを対象に，2 つの情報源共通のバーストの同定精度と，各情報源固有のバーストの同定精度を評価した結果を表 4 に示す．ここで，共通のバーストとは，ニュースとツイッターそれぞれで，同一トピックについて同日に同定されたバースト，固有のバーストとは，それ以外のバーストのことを指す．表より，共通のバースト，および，固有のバーストどちらについても精度よくバーストが同定できていることがわかる．唯一誤検出されたバーストは「政治」トピックであり，開会式前に両情報源で共通にバーストしている．これは，ロンドン五輪の開催期間において政治に関する文書が少なくなったことにより，開会式以前の期間がバースト期間と誤って同定されたものと考えられる．

さらに，共通のバーストにおいては，2 つの情報源間でバーストする原因は同じであった．このことから，高い精度で密接に関連するバーストを同定できていることがわかる．また，多数のトピックにおいて，ツイッター固有のバーストを数多く観測できていることがわかる．

次に，文書集合「ニュース記事 (ロンドン五輪)+ ツイート (ロンドン五輪)」から推定されたトピックである「レスリング」,「サッカー」,「選手の容姿」の 3 つのトピックに対するバーストの同定結果，および，トピックごとに 2 つの情報源間でバーストの対応を取った結果を図 1 ～ 図 3 に示す．まず，図 1 からは，ニュースとツイッターどちらについても，女子レスリングの小原選手，伊調選手，吉田選手が立て続けに金メダルを獲得した際にバーストしていることがわかる．一方で，男子レスリングの米満選手が金メダルを獲得した際には，ニュースでのみバーストしている．次に，図 2 からは，ニュースと比較して，ツイッターの方がより多くバーストしていることがわかる．ニュースでバーストしたのは，男子サッカー，女子サッカーの初戦が行われた時のみであるが，ツイッターでは，試合がある度にバーストしている．これは，サッカーの試合に対する，視聴者の関心の高さを表していると考えられる．最後の，図 3 は，選手の容姿に対する好意や感想を表すトピックであり，ツイッ

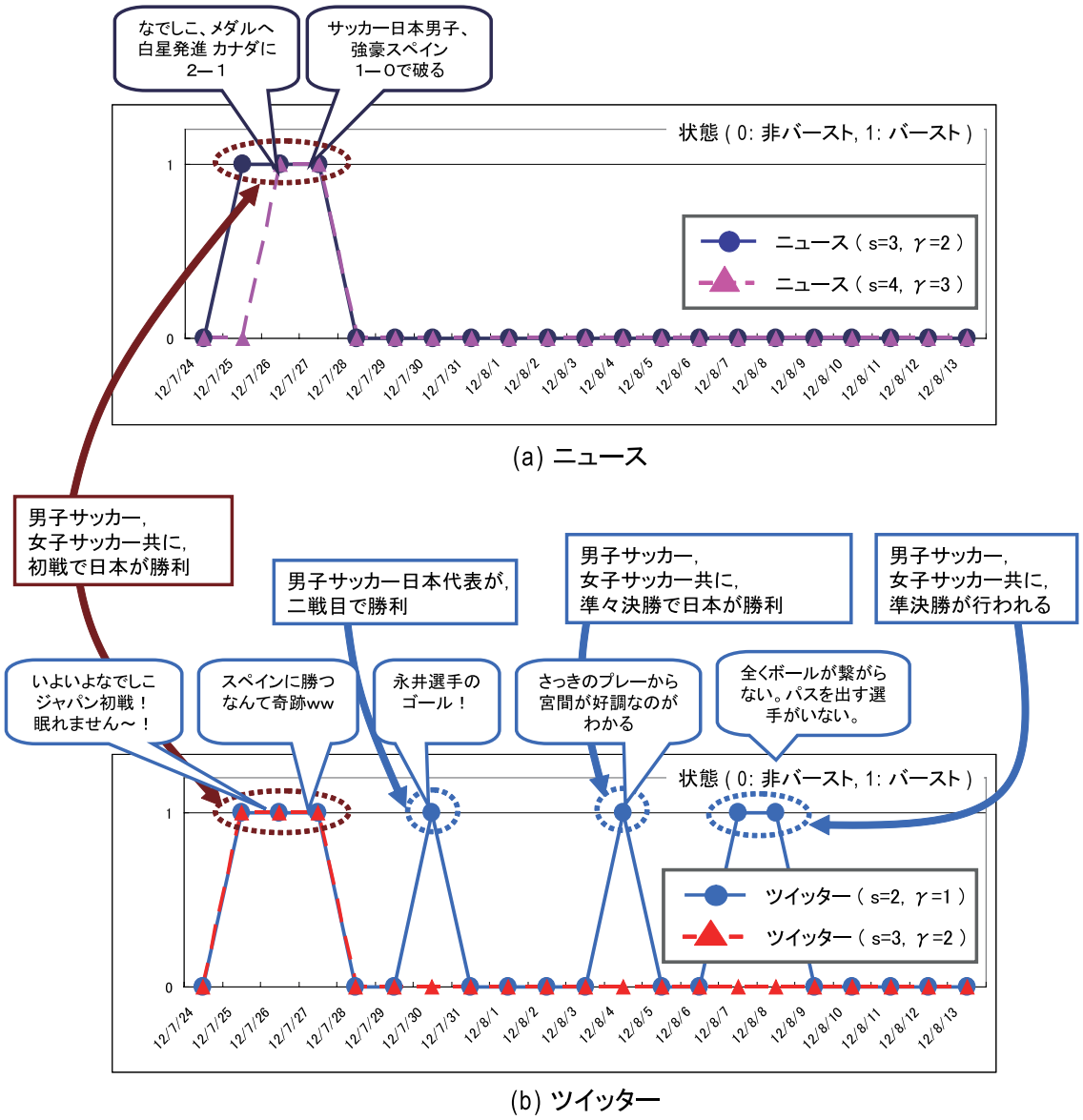



図 2　トピック「サッカー」におけるバーストの同定結果

ターでのみバーストするトピックである．この結果から，視聴者が，容姿について強く関心を抱いた選手を知ることができる．

## 8. 関 連 研 究

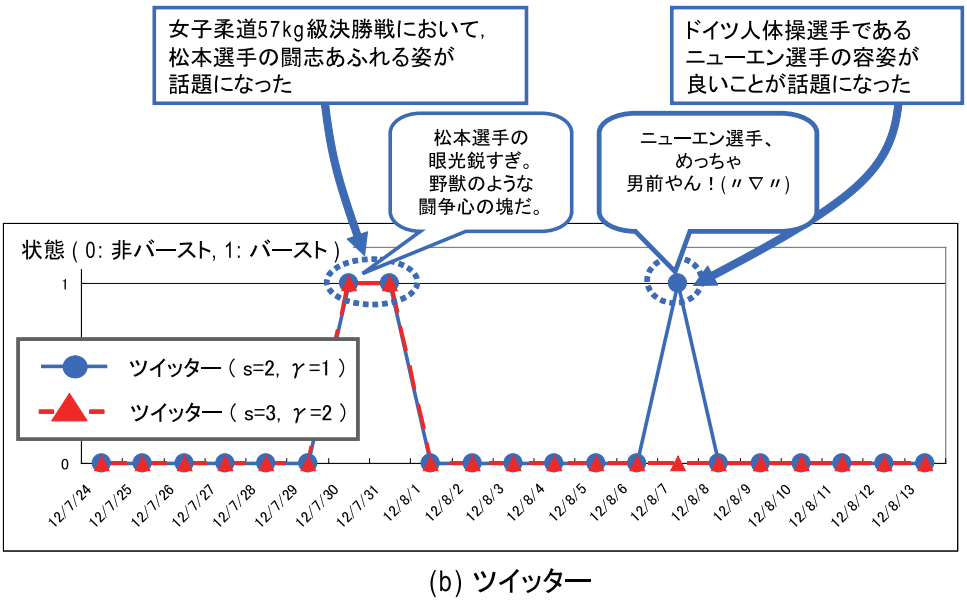
文献 [7], [8] においては，Kleinberg のバースト解析手法を用いて選定したバーストキーワードに対して，トピックへの集約を行う枠組みを提案している．しかし，これらは本研究とは異なり，DTM や LDA 等のトピックモデルを用いていない．文献 [7] では，共起度によってバーストキーワードを集約したものをトピックとし，トピックのバースト度やトピック間の関係性をグラフで視覚的に表示する手法を提案している．一方，文献 [8] ではバースト度の高い上位 20 キーワードを含む文書をクラスタリングし，その結果をもとに，話題ごとのキーワードの集約を行なっている．

本研究，並びに，文献 [6], [12] では，複数の情報源から生成される時系列文書集合に対してトピックを推定することにより，複数情報源間におけるトピックの対応付けを行っている．一方，文献 [4], [10] では，多言語の時系列ニュース記事を対象として，各言語のニュース記事集合に対してそれぞれ独立にトピックを推定した後，トピックを二言語間で対応付ける手法を提案している．このうち，文献 [4] においては，Wikipedia の言語間リンクを用いることにより二言語間の対応付けを行う．一方，文献 [10] では，トピックのバーストパターンの類似度を用いることにより二言語間の対応付けを行う．

なお，本研究に関連して，ニュース，ブログといった複数の相互に関連しあっている時系列の情報源を対象としてトピックモデルを適用し，各トピックの時系列の特徴をとらえる方式 [11] がある．

## 9. お わ り に

本論文では，DTM によって時系列文書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もとに，密接に関連しあう 2 種




(b) ツイッター

図 3　トピック「選手の容姿」におけるバーストの同定結果 (ツイッターでのみ観測)

類の情報源から生成される時系列文書集合から推定した時系列トピックを対象として，トピックのバーストを同定する方式を確立した．特に，この方式においては，情報源ごとに独立にトピックのバーストを同定する機能を実現した．時系列ニュース，および，ツイッターの 2 種類の情報源を対象として，この方式を適用し，その有用性を評価した結果，時系列ニュース，および，ツイッターの双方において，最大約 90% の精度でトピックのバーストを同定できることを実証した．また，時系列ニュースとツイッターの 2 つの情報源間で同一トピックについてのバーストが同日に発生した場合に，実際に 2 つのバーストが密接に関連する割合を評価した結果，87.5% と高い精度を達成した．一方，その他の多数のトピックにおいて，ツイートのみにおけるバーストを容易に観測することができた．

今後の課題は，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．

また，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] 等の考え方を導入することにより，本手法のオンライン化について検討する．

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

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

[4] S. Hu, Y. Takahashi, L. Zheng, T. Utsuro, M. Yoshioka, N. Kando, T. Fukuhara, H. Nakagawa, and Y. Kiyota. Cross-lingual topic alignment in time series Japanese / Chinese news. In *Proc. 26th PACLIC*, pp. 532–541, 2012.

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

[6] 小池大地, 牧田健作, 宇津呂武仁, 河田容英, 吉岡真治, 神門典子. 検索エンジン API を用いたウェブページ収集におけるトピックの多様性. 第 5 回 DEIM フォーラム論文集, 2013.

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

[8] 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 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2011.

[9] 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. 第 4 回 DEIM フォーラム論文集, 2012.

[10] X. Wang, CX. Zhai, and R. Sproat X. Hu. Mining correlated bursty topic patterns from coordinated text streams. In *Proc. 13th SIGKDD*, pp. 784–793, 2007.

[11] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In *Proc. 16th SIGKDD*, pp. 1079–10881, 2010.

[12] 鄭立儀, 胡碩, 小池大地, 宇津呂武仁, 吉岡真治, 神門典子. 時系列中国語ニュース・ブログにおけるトピックモデルの推定と比較対照分析. 言語処理学会第 19 回年次大会論文集, 2013.