Vol.2011-NL-204 No.6
2011/11/21

ニュースにおけるトピックのバースト特性の分析

高橋佑介[†1]　横本大輔[†1]
宇津呂武仁[†1]　吉岡真治[†2]

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バースト度を付与することで，トピック単位のバーストが検出可能であることを示す．

Analyzing Burst of Topics in News Stream

Yusuke Takahashi,[†1] Daisuke Yokomoto,[†1]
Takehito Utsuro[†1] and Masaharu Yoshioka [†2]

Among various types of recent information explosion, that in news stream is also a kind of serious problems. This paper studies issues regarding two types of modeling of information flow in news stream, namely, burst analysis and topic modeling. First, when one wants to detect a kind of topics that are paid much more attention than usual, it is usually necessary for him/her to carefully watch every article in news stream at every moment. In such a situation, it is well known in the field of time series analysis that Kleinberg's modeling of bursts is quite effective in detecting burst of keywords. Second, topic models such as LDA (latent Dirichlet allocation) and DTM (dynamic topic model) are also quite effective in estimating distribution of topics over a document collection such as articles in news stream. This paper focuses on the fact that Kleinberg's modeling of bursts is usually applied only to bursts of keywords but not to those of topics. Then, based on Kleinberg's modeling of bursts of keywords, we propose how to measure bursts of topics estimated by a topic model such as LDA and DTM.

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 によって分析期間におけるトピックの分布を推定したのち，推定されたトピックを対象としてバーストを検出する手法を提案する．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

†1 筑波大学大学院システム情報工学研究科　Graduate School of Systems and Information Engineering, University of Tsukuba
†2 北海道大学大学院 情報科学研究科　Graduate School of Information Science and Technology, Hokkaido University

Vol.2011-NL-204 No.6
2011/11/21

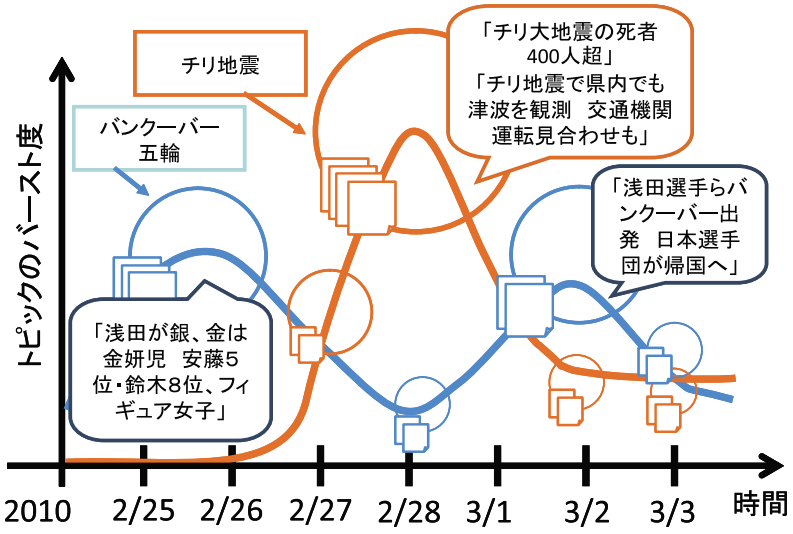


図 1　時系列ニュースにおけるトピックのバースト

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. バーストモデル

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[⋆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，分析期間全体を見たときの期待値 $p_0 = R/D$ 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．

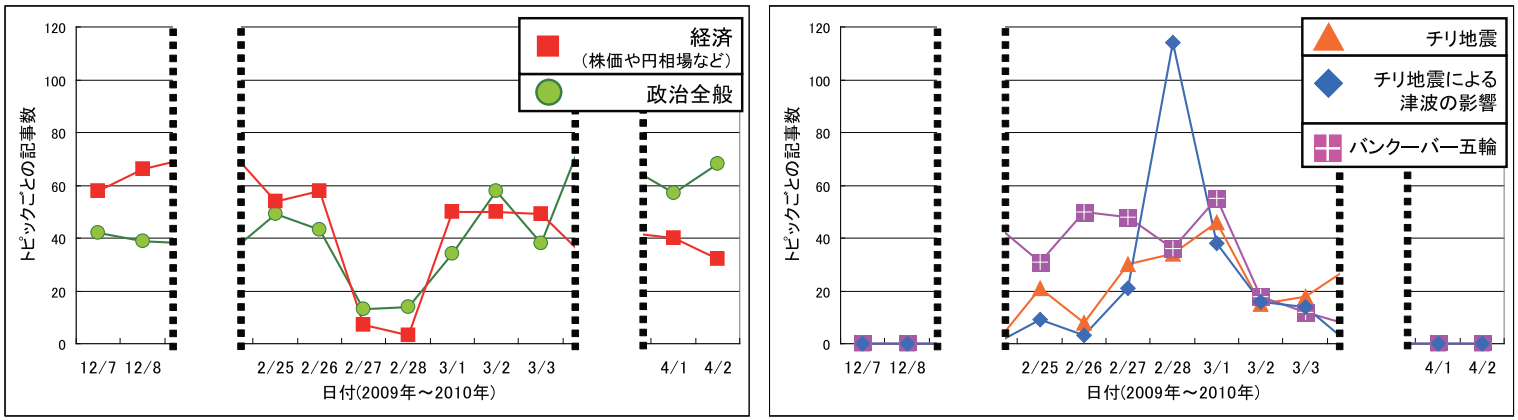
$$\sigma(i, r_t, d_t) = -\ln\left[\binom{d_t}{r_t} p_i^{r_t} (1 - p_i)^{d_t - r_t} \right]$$

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．

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

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．

⋆1 例えば，特定のキーワードを含む記事．



(a) 年間を通じて定常的に観測されるトピック

(b) 当該期間においてバーストしているトピック

図 2 主要トピックの記事数の遷移

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 日ごとにキーワードの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. トピックのモデル化

3.1 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)[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．

Vol.2011-NL-204 No.6
2011/11/21

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 α と，トピック数 K は，それぞれ $\alpha = 0.01$，$K = 20$ とした．

3.2 文書とトピックの対応付け

本研究では，一日ごとに，各トピックに対してニュース記事を一対一で割り当てることで，トピックごとのニュース記事集合の大きさを測ることとした．

ある日における文書集合を B，トピック数を K，1 つの文書を b $(b \in B)$ とすると，トピック z_n $(n = 1, \ldots, K)$ のニュース記事集合 $D(z_n)$ は以下の式で表される．

$$D(z_n) = \left\{ b \in B_t \mid z_n = \underset{z_u\ (u=1,\ldots,k)}{\operatorname{argmax}} p(z_u|b) \right\}$$

これはつまり，文書 b におけるトピックの分布において，確率が最大のトピックを文書 b に割り当てていることになる．

4. トピックへのバースト度の付与

本研究では，トピックのバーストを表現する手段として，各日における各トピックに対してバースト度を付与する．

t 番目の日付における各トピック z_n のバースト度 $bz(t, z_n)$ は，以下のように計算される．

$$bz(t, z_n) = \sum_w bw(t, w) \cdot p(w|z_n)$$

これは，各トピックから見たときの各キーワードのバースト度の重み付き総和に等しい．

しかし，本手法をそのままニュースに適用した場合，高頻度かつ周期的に出現するキーワードが，周期的に複数個バーストすることにより，例えば，株式市場に関するトピックなど，イベント性の少ないトピックのバースト度も高くなってしまう傾向があった．そこで今回は，年間を通して出現確率の高いキーワードのバースト度を 0 として扱うことにした．ここで，出現確率の高さの閾値は，予備実験の結果から 0.015 とした．

5. 分　　析

対象とした分析期間は，2009 年 12 月 7 日 ∼ 2010 年 4 月 2 日である．

ただしここで，バースト解析において用いる各キーワードの平均出現確率 p_0 の算出においては，2009 年 6 月 1 日 ∼ 2010 年 5 月 31 日の 1 年間のニュース記事集合[⋆2]を用いた．

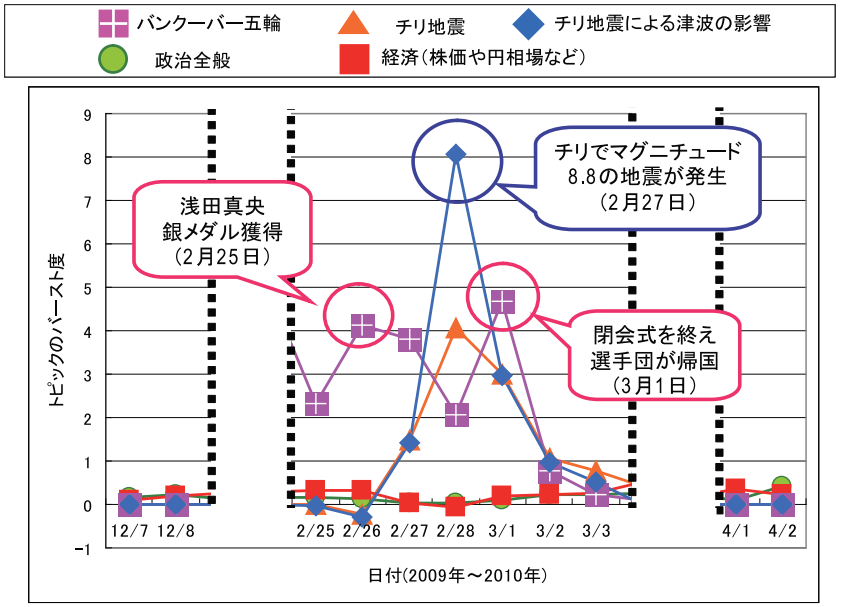


図 3 主要トピックのバースト度の遷移

はじめに，トピックごとの記事数の時間推移を表したグラフを図 2 に示す．(a) には年間を通じて定常的に観測されるトピック，(b) には当該期間においてバーストしているトピックを掲載した．ただし，12 月 8 日 ∼ 2 月 25 日，3 月 3 日 ∼ 4 月 1 日の期間については，簡単化のため省略した．ここで，トピックごとの記事数とは，3.2 節で定義したトピック z_n のニュース記事集合 $D(z_n)$ の要素数 $|D(z_n)|$ である．

図 2 (a) に示したトピックは，政治や経済という毎日定常的に現れるトピックであるため，年間を通じてほぼ同等の記事数である．対して，図 2 (b) には，2010 年 2 月 12∼ 28 日に開催されたバンクーバー五輪と，2010 年 2 月 27 日に発生したチリ地震という 2 つのトピックを示した．これらのトピックについては，年間を通じて継続して記事が観測される

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

⋆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 56,503 記事，38,758 記事，および，62,684 記事の合計 157,945 記事．

表 1 分析期間における主要なトピック

日付	バースト度上位 5 トピック 括弧内はトピックのバースト度	記事数上位 5 トピック 括弧内はトピックごとの記事数
2010 年 2 月 25 日	**トヨタリコール事件** (3.71), **バンクーバー五輪** (2.34), 芸能 (0.33), 経済 (0.33), 交通事故や地方のニュース (0.21)	企業 (63), 経済 (54), 製品情報 (51), 政治 (49), **トヨタリコール事件** (45)
2010 年 2 月 26 日	**バンクーバー五輪** (4.12), **トヨタリコール事件** (1.08), 学校に関する出来事 (0.36), 経済 (0.33), 芸能 (0.27)	企業 (77), 経済 (58), **バンクーバー五輪** (50), 政治 (43), 製品情報 (33)
2010 年 2 月 27 日	**バンクーバー五輪** (3.80), **チリ地震** (1.49), **チリ地震による津波の影響** (1.42), **トヨタリコール事件** (0.90), 学校に関する出来事 (0.25)	**バンクーバー五輪** (48), **チリ地震** (30), 北朝鮮問題 (23), 学校に関する出来事 (23), プロ野球 (22)
2010 年 2 月 28 日	**チリ地震による津波の影響** (8.07), **チリ地震** (4.06), **バンクーバー五輪** (2.06), 東京マラソン・気温 (0.25), プロ野球 (0.20)	**チリ地震による津波の影響** (114), **バンクーバー五輪** (36), **チリ地震** (34), プロ野球 (16), 製品情報 (15)
2010 年 3 月 1 日	**バンクーバー五輪** (4.67), **チリ地震** (3.00), **チリ地震による津波の影響** (2.96), **北教組幹部逮捕** (0.52), トヨタリコール事件 (0.46)	企業 (64), 製品情報 (55), **バンクーバー五輪** (22), 経済 (50), **チリ地震** (46)
2010 年 3 月 2 日	トヨタリコール事件 (1.20), **チリ地震** (1.08), **チリ地震による津波の影響** (0.97), **バンクーバー五輪** (0.78), 北教組幹部逮捕 (0.43)	政治 (58), 経済 (50), 企業 (45), 製品情報 (37), 交通事故や地方のニュース (32)
2010 年 3 月 3 日	**トヨタリコール事件** (1.91), **チリ地震** (0.77), **チリ地震による津波の影響** (0.52), **北教組幹部逮捕** (0.52), 経済 (0.27)	経済 (49), 企業 (46), 製品情報 (45), 刑事事件・裁判 (41), 政治 (38)

ということはなく，当該期間を中心とするわずかな期間でのみ記事が観測される

次に，本手法を用いてトピックに対してバースト度を付与し，トピックごとのバースト度の推移を表したグラフを図 3 に示す．ここで，図 3 上部の四角枠内はトピック名，及び，そのトピックのグラフのマーカーであり，吹き出し内には，そのトピックにおける主要な出来事と，その出来事が起こった日付を記した．

図 2 (a) で示したトピックと，図 2 (b) で示したトピックを，2 月 25 日 ∼ 3 月 3 日の期間における記事数で比較した場合，両者間で大きな差は見られない．しかし，図 3 で示したように，トピックに対してバースト度を付与することによって，特定の期間でのみ記事が観測されるトピックであるバンクーバー五輪とチリ地震を浮き彫りにすることができるようになった．これは，バーストの情報によって，単純な記事数とは異なる記事の時間軸方向の密度の変化が考慮されたためである．

ここで，DTM では，分析期間を通じてトピック数が一定であるという問題がある．そのため，本来ならチリ地震が起こっていない 2 月 27 日以前においても，チリ地震というトピックがあると見なされる．それにより，チリ地震と記述が似ているハイチ地震や南米について書かれた記事がチリ地震と対応付けられてしまい，図 2 のように，記事数を見ただけではチリ地震が発生した日を特定できない．しかし，図 3 に示すように，トピックにバースト度を付与することにより，チリ地震が発生した日を特定できることが分かる．

分析期間におけるその他の主要なトピックとして，2009 年 2 月 25 日 ∼ 3 月 3 日における，バースト度上位 5 トピック，及び，ニュース記事数上位 5 トピックを表 1 示す．この際，ある日におけるバースト度が 0.5 を超えるトピックについては太字で表した．

また，本手法を用いれば，トピックのバースト度に下限値を設けることによって，トピックのバーストを自動で検出することも可能である．一週間単位の期間を合計 5 期間選定し，バーストが検出されたトピックの個数とその正解数，及び，適合率を，バースト度の下限値ごとに表 2∼4 に示す．検出したトピック数の多さと適合率の高さを考慮すると，下限値が 0.5∼0.6 付近のとき精度よくバーストを検出することができることがわかる．

6. 関 連 研 究

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

表 2 バースト期間検出における適合率 (閾値 = 0.6)

期間	検出したトピック数	正解数	適合率 [%]
2009 年 7 月 5 日 ～ 11 日	11	11	100
2009 年 9 月 20 日 ～ 9 月 26 日	8	8	100
2009 年 10 月 5 日 ～ 10 月 11 日	6	6	100
2010 年 1 月 9 日 ～ 1 月 15 日	18	17	94.4
2010 年 2 月 25 日 ～ 3 月 3 日	21	21	100

表 3 バースト期間検出における適合率 (閾値 = 0.5)

期間	検出したトピック数	正解数	適合率 [%]
2009 年 7 月 5 日 ～ 11 日	15	12	80.0
2009 年 9 月 20 日 ～ 9 月 26 日	9	9	100
2009 年 10 月 5 日 ～ 10 月 11 日	9	8	88.9
2010 年 1 月 9 日 ～ 1 月 15 日	19	18	94.7
2010 年 2 月 25 日 ～ 3 月 3 日	24	24	100

表 4 バースト期間検出における適合率 (閾値 = 0.4)

期間	検出したトピック数	正解数	適合率 [%]
2009 年 7 月 5 日 ～ 11 日	17	12	70.6
2009 年 9 月 20 日 ～ 9 月 26 日	15	14	93.3
2009 年 10 月 5 日 ～ 10 月 11 日	10	8	80.0
2010 年 1 月 9 日 ～ 1 月 15 日	24	18	75.0
2010 年 2 月 25 日 ～ 3 月 3 日	26	26	100

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，トピックモデルとして DTM を用いることにより，時系列ニュースにおける各日のニュース記事集合におけるトピック分布を推定し，キーワードのバーストの考え方を拡張することでトピックに対してバースト度を付与する手法を提案した．具体的には，トピックごとの各キーワードの条件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ことで算出した．

これにより，キーワードに比べてより情報量の大きな，トピックという単位でバーストを捉えることができるようになり，各トピックのバースト度，記事数における時間推移を見ることでトピックの特徴や，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．

本論文では，キーワードに対してバースト解析を行ってバースト度を付与した後，キーワードのバースト度をトピックモデルによって推定したトピックに適用した．一方，今後は，それとは異なり，トピックモデルによって推定したトピックの単位を対象として，直接バースト解析を行いバースト度を測定する方式を定式化し，本論文の手法との比較を行う．また，本論文では，トピックモデルとして DTM を適用したが，その他のトピックモデルとして，On-line LDA[1] によって推定されたトピックを対象として本論文の手法を適用する予定である．

参 考 文 献

1) ALSumait, L., Bardara, D. and Domeniconi, C.: On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking, *Proc. 8th ICDM*, pp.3–12 (2008).
2) ALSumait, L., Bardara, D., Gentle, J. and Domeniconi, C.: Topic Significance Ranking of LDA Generative Models, *Proc. ECML/PKDD*, pp.67–82 (2009).
3) Blei, D.M. and Lafferty, J.D.: Dynamic Topic Models, *Proc. 23rd ICML*, pp.113–120 (2006).
4) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
5) Kleinberg, J.: Bursty and Hierarchical Structure in Streams, *Proc. 8th SIGKDD*, pp.91–101 (2002).
6) Mane, K. and Borner, K.: Mapping topics and topic bursts in PNAS, *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).