Vol.2010-MPS-81 No.10
Vol.2010-BIO-23 No.10
2010/12/16

文書ストリームからのバースト潜在トピック抽出における t-LDA 法の性能検証

水田昌孝[†1]　熊野雅仁[†2]
小野景子[†2]　木村昌弘[†2]

我々は以前に，文書ストリームからバースト潜在トピック抽出する t-LDA 法を提案した．t-LDA 法は，潜在トピックを抽出するために文書生成確率モデル LDA(Latent Dirichlet Allocation) を用い，バーストトピックを同定するために時間フィルタを導入している．そして，LDA と時間フィルタに基づいて，時間情報を持つ 2 つの文書間の類似度を構築し，階層的クラスタリング法を適用することで文書ストリームからバースト潜在トピックを抽出している．本稿では，人工データを用いた実験により t-LDA 法の定量的な有効性を検証し，オンラインニュースデータを用いた実験により t-LDA 法の有効性を実証する．

Performance Verification of t-LDA Method for Extracting Bursty Latent Topics from a Document Stream

Masataka Mizuta,[†1] Masatoshi Kumano,[†2]
Keiko Ono[†2] and Masahiro Kimura[†2]

We previously proposed the t-LDA method that extracts bursty latent topics from a documet stream. The method utilizes Latent Dirichlet Allocation (LDA), which is a probabilistic generative model of documents, for extracting latent topics, and introduce a time-filter for identifying bursty topics. It constructs a measure of similarity between two documents with time-stamps on the basis of LDA and the time-filter, and extract bursty latent topics from a document stream by applying a hierarchical agglomerative clustering method. In this paper, we quantitatively verify its effectiveness by using synthetic data, and demonstrate its effectiveness by using real online news data.

1. はじめに

近年，Web 上ではニュース文書やブログ，電子メールといった様々な種類の文書ストリームが存在するようになった．しかし，これらの文書ストリームは，多くの人から発信される情報を刻々と蓄積し続けるため，その情報量が膨大となり，全体像を把握する事が困難となっている．そのため，このような膨大な情報量を持つ文書ストリームに対して，自動的に文書の情報を分析する技術の重要性が高まっている．その技術の一環として，文書が持つトピックに着目した研究が盛んに行われている．

データマイニングによって文書ストリームからトピックを抽出する関連研究としては，文書ストリームからバースト性の高いトピックとその期間を抽出することで，文書ストリームを構成する文書群の主なトピックを把握する手法というが Kleinberg[1] によって報告されている．ただし，この手法におけるトピック抽出は，各トピックを象徴するような特定の単語の出現頻度に基づいて行われるため，各トピックを明示的に特徴づける単語に依存しないような，潜在的なトピックの抽出を行うことは難しい．また，トピックは，各文書に単独で存在する場合よりも，複数のトピックが混在する場合が多いため，多重性を考慮したトピック抽出を行うことが望まれる．そこで，Blei ら [2] によって報告された教師なし学習を可能とする多重トピックモデル LDA(Latent Dirichlet Allocation) により，文書が潜在的に有する多重トピックを推定するという手法が注目されている．そこで，我々は LDA を利用して推定した潜在的な多重トピックと，文書の生成時間情報に基づく時間フィルタを導入する事で，文書ストリームから潜在的な多重トピックの類似性が高く，かつバースト性も有するような文書群をクラスタリングし，バースト潜在トピックを抽出するという手法を報告した [3]．本稿では，この手法を便宜上 t-LDA 法と呼ぶ．この t-LDA 法によるバースト潜在トピック抽出は，毎日新聞の文書データを用いた実験によって，有効性を示している．

本研究では，t-LDA 法がバースト性を有する潜在トピックに関して，実際にどのような性能を示すかを検証するため，人工データを用いた実験により，定量的に評価を行う．また，毎日新聞のような単独のメディアに関する実験ではなく，多数のメディアから文書ストリームが流入する Google ニュースを対象とすることで，より多様性があると思われる文書スト

†1 龍谷大学大学院 理工学研究科 電子情報学専攻
Division of Electronics and Informatics, Ryukoku University
†2 龍谷大学 理工学部 電子情報学科
Department of Electronics and Informatics, Ryukoku University

Vol.2010-MPS-81 No.10
Vol.2010-BIO-23 No.10
2010/12/16

リームに対しても実験を行う．

2. バースト潜在トピック抽出問題

本研究では，文書ストリーム $D = \{d_{t,m}; t = 1, ...T, m = 1, ...M\}$ における，バースト潜在トピック文書群 $D_l \subset D$ $(l = 1, \cdots, L)$ とそのバースト期間 $[T_{l,0}, T_{l,1}]$ $(l = 1, \cdots, L)$ を抽出する，バースト潜在トピック抽出問題を扱う[3)]．ここに，各文書 $d_{t,m}$ は BOW(bag-of-words) 表現されている．

3. t-LDA 法

t-LDA 法では，BOW 表現された文書ストリームから LDA を用いて推定したトピックベクトルと，文書の生成時間情報に基づく時間フィルタによって文書間距離 S を求める．そして，この文書間距離に基づいて構成したデンドログラムからクラスタリングを行うことによって，バースト潜在トピックの抽出を行う．この一連の流れを図 1 に示す．なお，トピックの次元数 k，時間フィルタにおいて，文書生成時間に基づく類似度が最大である期間 τ_1，文書生成時間に基づく類似度が 0 となる期間 τ_2，バースト性の有無を判定する閾値 I，およびノード間の最大文書生成時間差 J は，それぞれ任意に指定可能なパラメータである．

3.1 LDA を用いたトピックベクトルの抽出

LDA における 1 文書の生成過程を以下に述べる．

Step1. ディリクレパラメータ$\boldsymbol{\alpha}$からトピックベクトル$\boldsymbol{\theta}$を求める

Step2. 以下を文書の単語総数 N 回だけ反復

Step2.1. トピックベクトル$\boldsymbol{\theta}$からトピック z を選択

Step2.2. トピック z と単語生成確率ベクトル$\boldsymbol{\beta}$から単語 w を 1 つ選択

この生成過程を M 回反復することによって，M 文書からなる文書ストリームを得ることができる．また，1 文書における生成過程を数式で表現すると，

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{i=1}^{N} p(z_i|\boldsymbol{\theta}) p(w_i|z_i, \boldsymbol{\beta}) \tag{1}$$

となる．また，1 文書における事後確率の周辺分布は次式によって表される．

$$p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \Big(\prod_{i=1}^{N} \sum_{z_i} p(z_i|\boldsymbol{\theta}) p(w_i|z_i, \boldsymbol{\beta}) \Big) d\boldsymbol{\theta} \tag{2}$$

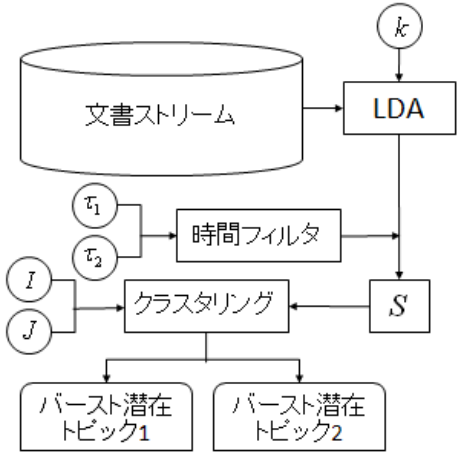

図 1 t-LDA 法によるバースト潜在トピック抽出のフロー
Fig. 1 The flow for extracting bursty latent topics by the t-LDA method

式 (2) のパラメータを求める事で，各文書のトピックベクトル$\boldsymbol{\theta}$を求めることが出来るが，このトピックベクトル$\boldsymbol{\theta}$の次元が大きい場合，各パラメータを解析的に求める事が困難となる．そこで，本研究では，式 (3) で表す変分事後分布を導入し，EM アルゴリズムを用いて近似的に各パラメータを求める．

$$q(\boldsymbol{\theta}, z|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{i=1}^{N} p(z_i|\boldsymbol{p}hi) \tag{3}$$

なお，$\boldsymbol{\gamma}$，$\boldsymbol{\phi}$はそれぞれディリクレパラメータ，多項パラメータを表す．本研究では，このパラメータ$\boldsymbol{\gamma}$を各文書の持つトピックベクトルを近似したものとして用いる．

3.2 文書間距離の定義

文書間距離の定義には，LDA により推定した各文書のトピックベクトル$\boldsymbol{\gamma}$の他に，文書生成時間情報も特徴量として用いる．文書 m と文書 n の生成時間差 $t_{m,n}$ に基づく時間フィルタ $T_{m,n}$ を次式に示す．

Vol.2010-MPS-81 No.10
Vol.2010-BIO-23 No.10
2010/12/16

$$T_{m,n} = \begin{cases} 1 & (t_{m,n} \leq \tau_1) \\ 1 - \frac{t_{m,n}-\tau_1}{\tau_1-\tau_2} & (\tau_1 \leq t_{m,n} \leq \tau_2) \\ 0 & (\tau_2 \leq t_{m,n}) \end{cases} \quad (4)$$

なお，パラメータ τ_1 は文書生成時間に基づく類似度が最大値を保つフィルタなし期間，τ_2 は文書生成時間に基づく類似度が 0 となるフィルタ限界期間である．文書 m と文書 n の文書生成時間差に影響する時間フィルタ $T_{m,n}$ とトピックベクトル$\boldsymbol{\gamma}$から，文書 m と文書 n の文書間距離を次式で定義する．

$$S_{m,n} = 1 - \cos(\boldsymbol{\gamma}_m, \boldsymbol{\gamma}_n) * T_{m,n} \quad (5)$$

ここで，パラメータ $\boldsymbol{\gamma}_m$ は文書 m におけるトピックベクトルを表し，パラメータ $\boldsymbol{\gamma}_n$ は文書 n のトピックベクトルを表す．$S_{m,n}$ は値が小さいほど文書 m と文書 n の潜在的なトピックベクトル，そして文書生成時間における類似性が高くなることを示す．

3.3 バースト潜在トピック抽出

文書間距離 S に基づいて，群平均法による階層的クラスタリングを行う．ここで，ある階層のクラスタ内におけるノード間の最大生成時間差 J に基づいてクラスタ数を決定する．クラスタの統合を行う場合，統合段階でクラスタ内の全ての J を算出し，J が閾値を越えた場合，その統合を行わないという方式を採用する．これにより，下位からクラスタの統合が開始され，J に基づいてクラスタ数が自動的に決定できる．

このようにして決定したクラスタ内の文書群は，生成された時期が近く，かつ文書間距離が近いという特徴をもつ．ここで，抽出されたクラスタ内の文書群は，最大生成時間差 J 以内に収まる関係を持つため，その期間内に含まれる文書数が多いほど，短期間に類似性の高い文書が高頻度で現れるというバースト性を表現しているものと考えられる．そこで，ノード数がバースト基準となる閾値 I 以上である場合，そのクラスタはバースト性を持つものと考える．以上より，本研究では，バースト性を持つと判定されたクラスタに属する文書群を，ある粒度のバースト潜在トピックとする．

4. 性能評価と比較法

本稿では，t-LDA 法によるバースト潜在トピック抽出実験の性能を相対的に評価するために，比較法として文書間距離の定義を変更した次の 2 つの手法を行う．

[比較法 1]

比較法 1 では，時間フィルタを用いず，文書のトピックベクトル$\boldsymbol{\gamma}$ のみを特徴量文書間距離 S' を次式で定義する．

$$S'_{m,n} = 1 - \cos(\boldsymbol{\gamma}_m, \boldsymbol{\gamma}_n) \quad (6)$$

この文書間距離によってバースト潜在トピック抽出を行う手法を，便宜上 LDA 法と呼ぶ．

[比較法 2]

比較法 2 では，tfidf によって重み付けされた単語ベクトル $\mathrm{w} = \{w_i : i = 1, \ldots, N\}$ と，t-LDA 法に用いた文書生成時間に基づく時間フィルタを特徴量として用いる．この w は，次式によって重み付けを行う．

$$\begin{aligned} w_i &= tf_i * idf_i \\ tf_i &= \frac{w_i}{\sum_k w_k} \\ idf_i &= \log M / |d : c_i \in d| \end{aligned} \quad (7)$$

$|d : d \ni w_i|$ は，単語 w_i を含む文書数である．上式により重み付けされた単語ベクトル$\boldsymbol{w}$ より，文書間距離を次式のように定義する．

$$S''_{m,n} = 1 - \cos(\boldsymbol{w}_m, \boldsymbol{w}_n) * T_{m,n} \quad (8)$$

このように定義された文書間距離に基づくバースト潜在トピック抽出を，本稿では tf-idf 法と呼ぶ．

5. 人工データによる実験

この実験では，人工データを用いることで，各手法におけるバースト潜在トピック抽出の性能を定量的に評価し，有効性を明らかにすることを目的としている，

人工データを生成する際，文書ストリーム中に出現するバースト潜在トピックの種類数，各バースト潜在トピックを構成する文書群$\boldsymbol{d}$およびトピックベクトル$\boldsymbol{\gamma}$，そして単語生成確率ベクトル$\boldsymbol{\beta}$を既知のデータとして用いる．各文書の生成手法は 3.1 章に掲載した LDA の文書生成過程に倣い，トピックベクトル$\boldsymbol{\gamma}$から選択されたトピック z と単語生成確率ベクトル$\boldsymbol{\beta}$か 1 つの単語を選び，これを N 回繰り返すことで，総数 N の単語からなる BOW 表現された文書を 1 つ生成するという方式を採用している．

また，バースト潜在トピックを構成する文書群は，一定期間中に集中して出現するものとし，また，どのバースト潜在トピックとも関連性を持たない文書は，文書ストリームの全期間にわたって出現するものとする．このとき，各バースト潜在ピックが発生した期間 t と生成された文書群$\boldsymbol{d}$を，バースト潜在トピックにおける真のデータとして用いる．

このように生成された文書ストリームを用いて，t-LDA 法によるバースト潜在トピック抽

Vol.2010-MPS-81 No.10
Vol.2010-BIO-23 No.10
2010/12/16

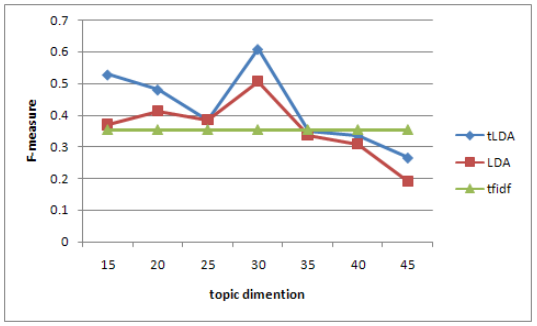

図 2 人工データを用いたバースト潜在トピック抽出結果
Fig. 2 Resluts for extracting bursty latent topics using the syntheticl data

出実験を行う．ここで，真のバースト潜在トピック$\boldsymbol{B}_d$と，t-LDA 法により抽出されたバースト潜在トピック$\boldsymbol{B}_d$' の一致度を，Presion と Recall による調和平均で求められる F 値によって，定量的に評価する．

$$P = \frac{\sum \boldsymbol{B}_d \boldsymbol{B}_d{}'}{\sum \boldsymbol{B}_d} \quad , \quad R = \frac{\sum \boldsymbol{B}_d \boldsymbol{B}_d{}'}{\sum \boldsymbol{B}_d{}'}$$
$$F = \frac{2PR}{P+R} \tag{9}$$

5.1 実験データ

実験では，トピック次元数 k を 30，単語種類数 V を 10000 とした単語生成確率ベクトル$\boldsymbol{\beta}_{k,V}$ から 2000 の文書を生成した．なお，単語生成確率ベクトル$\boldsymbol{\beta}$は，各トピックを明示的に特徴づける単語頻度が約 350 語含まれるように設定した．また，文書ストリームの総期間を 200，時間フィルタにおけるフィルタなし期間 τ_1 を 7，フィルタ限界期間 τ_2 を 14 と指定した．そして，ノード間最大生成時間差 J は τ_2 と同じ 14 とし，バースト性の有無を判定する基準値 I は 10 と定めた．真のバースト潜在トピック数は 15 とし，各バースト潜在トピックが生じる期間は時間フィルタのフィルタ限界期間 τ_2 と等しい 7 とした．

なお，t-LDA 法と LDA 法はバースト潜在トピック抽出を行う際，トピックベクトル$\boldsymbol{\gamma}$の推定のためにトピックベクトルの次元 k を指定する必要があるが，ここでは k を未知のパラメータとして扱い，k= $\{15, 20, 25, 30, 35, 40, 45\}$ と変更して，それぞれ実験を行った．

表 1 人工データに対して抽出されたバースト期間 (k = 30)
Table 1 The bursty spans extracted for the synthetic data (k = 30)

ID	t-LDA 法	LDA 法	tfidf 法
1	9(174〜188)	183(15〜198)	5(9〜14)
2	6(56〜62)	112(15〜127)	4(163〜167)
3	9(25〜34)	172(17〜189)	11(31〜42)
4	11(128〜139)	193(6〜199)	7(172〜179)
5	10(58〜68)	117(21〜138)	8(131〜138)
6	6(71〜77)	156(30〜186)	7(59〜66)
7	14(1〜15)	172(1〜173)	7(92〜99)
8	6(26〜32)	147(33〜180)	6(99〜105)
9	13(61〜74)	116(23〜139)	14(43〜57)
10	5(99〜104)	185(1〜186)	13(27〜41)
平均生成時間差	8.9	155.3	8.2

5.2 結　　果

人工データを対象としたバースト潜在トピック抽出結果を図 2 に示す．ここで，各手法における F 値の平均は，t-LDA 法が 0.42，LDA 法が 0.36，tfidf 法が 0.35 となった．F 値が最大となった点はトピックベクトルの次元を k=30 としたときの t-LDA 法による抽出結果であり，このとき，F 値は 0.61 という値を示した．

しかし，k が 30 よりも大きくなったとき，t-LDA 法と LDA 法による抽出精度は急速に落ち込み，k=35 の時点からは k の変動による影響を受けない tfidf 法よりも F 値が下回るという結果になった．

この結果から，トピックベクトル次元数 k を真の値に近づけることができると，t-LDA 法はこの 3 つの手法の中で，文書ストリーム中のバースト潜在トピック抽出に最も適した手法ということができる．しかし，k が真の値より大きくなる場合，LDA により推定したトピックベクトルを特徴量として用いる手法は性能が低下することが判明した．

次に，バースト期間に焦点を当てる．表 1 に，k=30 のときに抽出されたバースト潜在トピックにおける，バースト期間を示す．時間フィルタを用いる t-LDA 法や tfidf 法に比べ，バースト潜在トピック抽出に文書生成時間を用いない LDA 法では，文書が生成される期間の幅が突出して大きい．そのため，t-LDA 法や tfidf 法では別々のバースト潜在トピックとして抽出されるものでも，LDA 法を用いると同一のバースト潜在トピックと判定されてしまう場合がみられた．この結果より，膨大な文書ストリーム中で生じている事象を細かく分析する場合等で，時間フィルタを類似度として用いる有効性が示された．

Vol.2010-MPS-81 No.10
Vol.2010-BIO-23 No.10
2010/12/16

6. オンラインニュースデータによる実験

この実験では，多数のメディアから文書が流入するような文書ストリームに対しても，バースト潜在トピック抽出が有効性を示すか検証するために，実際のオンラインニュースデータを用いた実験を行う．使用したデータと，t-LDA 法によるバースト潜在トピック抽出によって得られた実験結果を以下に述べる．

6.1 実験データ

本研究では，多様性を持つと思われる文書ストリームとして，Google ニュース社会カテゴリの文書を実験データとして使用した．この文書ストリームは 2010 年 6 月 8 日から 8 月 8 日までの二ヶ月間に渡って収集したものであり，文書ストリームに含まれる文書の総数は 5434, 単語種類の総数は 16274 であった．また，時間フィルタにおけるフィルタなし期間 τ_1 は文書ストリーム総期間 60 日の 3 %である 2 日，フィルタ限界期間 τ_2 は τ_1 の 2 倍である 4 日といったように割り当てた．ノード間の最大生成時間差 J は τ_2 と同じ 4 日とし，バースト性の有無を判定する閾値 I は 10 とした．つまり，同一，あるいは類似性の高い多重トピックを持つ文書が，3 日の間に 10 以上出現している場合，これらの文書群はバースト性を有するノードと判定し．抽出を行うことを意味している．

6.2 結　　果

t-LDA 法におけるバースト潜在トピック抽出結果を表 2 に示す．このとき，トピックベクトルの次元 k を 30 として各文書のトピックベクトルを推定し，文書間距離を求める特徴量として用いている．表中の「バースト潜在トピック」は，抽出された文書群の内容を反映すると思われるアノテーションを人手で付与したものである．また，表 2 では，抽出されたバースト潜在トピックのうち，含有ノード (文書) 数の大きい上位 10 種類を報告している．

また，図 3 に t-LDA 法で抽出された実際に文書群の一例を示す．これは表 1 における ID 8，バースト潜在トピック「夏のイベント関連」を構成する文書群の文書生成時間とタイトルである.「土用の丑の日」といった単語のように，tfidf 法で見られるような特徴的な単語の出現頻度に類似性が見られる文書の他に，地方のイベントといった単語の出現頻度という観点からは類似性が見られる文書だけでなく，文書内容において類似性の見られる文書も同一トピックとして分類されていることがわかる．この結果は，共通して夏に開催されるイベントを同一トピックにクラスタリングし，抽出するということを実現させた事を示している．

次に，同一のバースト潜在トピックに含まれるような文書群を，時間フィルタによって分離し，抽出している例を図 4 と図 5 に示す．これは表 2 における ID1，ID3 の「集中豪雨関連」を構成する文書の一部である．これらの文書は，文書中に出現する単語においても文書の内容においても類似性が高いように思われるが，文書の生成された時間に差があるため，時間フィルタによって別々のバースト潜在トピックに分類されている．この結果から，同じ記録的な雨に関する文書でも細かい分析を可能としていることがわかる．

以上の実験結果から，t-LDA 法によるバースト潜在トピック抽出は，Google ニュースのような多数のメディアから文書が流入する文書ストリームに対しても，有効性を示すと考えることができる．

表 2　オンラインニュースデータを用いたバースト潜在トピック抽出結果
Table 2　Results for extracting bursty latent topics using the online news data

ID	バースト潜在トピック	ノード数	バースト期間
1	集中豪雨関連 1	33	7/11〜7/14
2	原爆関連	15	8/4〜8/7
3	集中豪雨関連 2	15	6/20〜6/23
4	お祭り関連	14	8/1〜8/3
5	阿久根市長関連 1	14	8/4〜8/6
6	国会関連	12	7/20〜7/21
7	参院選関連 1	12	7/9〜7/11
8	夏のイベント関連 1	11	7/25〜7/26
9	ガス田交渉関連	11	7/26〜7/27
10	猛暑関連	10	7/26〜7/27

- 7/25 2:30:53 夏祭りの山車，見物客はねる＝４人重軽傷
- 7/25 15:33:42 中津祇園山車が衝突事故中津市の夏祭り「中津祇園」
- 7/25 16:58:37 みんなでごしごし『日本橋』保存会が橋洗い
- 7/25 18:59:37 北上川ゴムボート川下り大会：「流れ速く疲れた」
- 7/25 19:2:22 夏祭りの山車，見物客はねる…４人重軽傷大分県中津市の夏祭り
- 7/25 19:5:30 猛暑はウナギで乗り切れ！「土用の丑」で百貨店もぎわう
- 7/25 20:27:11 土用丑の日うなぎ店大忙し 26 日は土用の丑の日
- 7/25 23:13:5 土用の丑の日：食欲そそる香り２６日は「土用の丑（うし）の日」
- 7/26 1:48:18 「土用の丑」暑いよ熱いよ大忙し猛暑が続く東海地方

図 3　オンラインニュースデータから t-LDA 法によって抽出されたバースト潜在トピック (ID 8)
Fig. 3　The bursty latent topic (ID 8) extracted by t-LDA method from the online news

Vol.2010-MPS-81 No.10
Vol.2010-BIO-23 No.10
2010/12/16

・7/13 3:17:13 九州から関東，大雨続く
・7/13 3:17:13 九州から関東，大雨続く
・7/13 6:20:44 九州各地で大雨被害１４日も
・7/13 19:24:18 大雨：九州から東北各地に大雨
・7/14 0:55:29 大雨：西日本で記録的大雨
・7/14 4:58:47 九州～東北激しい雨のおそれ
・7/14 14:1:14 九州から東北にかけて大雨警戒呼びかけ
・7/14 15:27:16 大雨：九州から東北で大雨に警戒を
・7/14 18:44:15 山口県で非常に激しい雨梅

図 4 オンラインニュースデータから t-LDA 法によって抽出されたバースト潜在トピック (ID 1)
Fig. 4 The bursty latent topic (ID 1) extracted by t-LDA method from the online news

・6/20 5:25:7 鹿児島で記録的大雨，土砂災害に警戒
・6/20 5:41:25 九州南部に大雨，警戒を＝梅雨前線が活発
・6/20 6:35:11 九州南部で豪雨，空や鉄道混乱…
・6/20 8:24:37 九州南部激しい雨 鹿児島では避難勧告も新幹線運休相次ぐ
・6/20 22:17:33 九州南部に大雨続く，新幹線４本が運休
・6/21 8:46:50 鹿児島県本土，なお大雨の恐れ
・6/21 16:40:56 九州で再び激しい雨の恐れ
・6/22 0:0:59 大雨：厳重注意あす朝まで前線が停滞
・6/22 2:51:7 九州南部，大雨に警戒＝土砂災害の恐れ高まる
・6/22 6:52:12 九州南部土砂災害に厳重警戒

図 5 オンラインニュースデータから t-LDA 法によって抽出されたバースト潜在トピック (ID 1)
Fig. 5 The bursty latent topic (ID 1) extracted by t-LDA method from the online news

7. ま と め

文書ストリームからバースト潜在トピックを抽出する手法である t-LDA 法の性能を評価した．まず，人工データを用いた実験によりその有効性を定量的に確認した．次に，オンラインニュースデータである Google ニュースを用いた実験により，その有効性を実証した．

参 考 文 献

1) Kleinberg. J: *Bursty and Hierarchical structure in streams &* , Proceedings of the 8th ACM SIGKDD International & Conference on Knowledge Discovery and Data Mining (KDD-03), pp. 91-101 (2002).
2) D. M. Blei, A. Y. Ng, M. I. Jordan:*Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 pp.993-1002, (2003).
3) 水田昌孝, 熊野雅仁, 木村昌弘：LDA と時間フィルタを用いた文書ストリームからのバースト潜在トピック抽出，人工知能学会 KBS 研究会, 2010.