情報処理学会論文誌　数理モデル化と応用　Vol.7　No.1　53–60 (Mar. 2014)

# Twitterにおけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案

佐々木 謙太朗[1,a)]　吉川 大弘[1]　古橋 武[1]

受付日 2013年8月26日, 採録日 2013年9月27日

**概要**：Latent Dirichlet Allocation（LDA）は，様々な分野で応用されているトピックモデルであり，Twitterにおけるユーザ属性の推定や話題の要約などに適用した研究も数多く報告され始めている．LDAをツイート集合に適用する場合，1ツイートを1文書とすると，文書の短さやノイズの多さにより，LDAが有効に機能しないことが多いため，1ユーザの全ツイートを1文書とする方法が一般的に用いられる．これに対して，1ツイートが1トピックからなるという仮定に基づいたトピックモデルであるTwitter-LDAが提案され，前者の方法に比べて，トピックの意味のまとまりの面で優れていると報告されている．しかし一方でTwitter-LDAは，オンライン学習ができないという課題がある．本論文では，Twitter-LDAを改良し，Twitterに適したオンライン学習可能なトピックモデルを提案する．提案モデルでは以下の2点についてTwitter-LDAを拡張する．第1に，一般語とトピック語との比率をユーザごとに推定することで，より高精度にツイートの生成過程をモデル化する．第2に，ユーザの購買行動をモデル化したTopic Tracking Model（TTM）の機構をモデルに加えることで，Twitterにおけるユーザの興味と話題の時間発展をオンラインで学習可能とする．

キーワード：トピックモデル，Twitter，時間発展，オンライン学習

# A Proposal of Online Topic Model for Twitter Considering Temporal Dynamics of User Interests and Topic Trends

Kentaro Sasaki[1,a)]　Tomohiro Yoshikawa[1]　Takeshi Furuhashi[1]

Received: August 26, 2013, Accepted: September 27, 2013

***Abstract:*** Latent Dirichlet Allocation (LDA) is a topic model which has been applied to various fields. It has been also applied to user profiling or event summarization on Twitter. In the application of LDA to tweet collection, it generally treats aggregated all tweets of a user as a single document. On the other hand, Twitter-LDA which assumes a single tweet consists of a single topic has been proposed and showed that it is superior to the former way in topic semantic coherence. However, Twitter-LDA has a problem that it is not capable of online inference. In this paper, we extend Twitter-LDA in the following two points. First, we model the generation process of tweets more accurately by estimating the ratio between topic words and general words for each user. Second, we enable it to estimate temporal dynamics of user interests and topic trends in online based on Topic Tracking Model (TTM) which models consumer purchase behaviors.

***Keywords:*** topic model, Twitter, time evolution, online learning

## 1. はじめに

近年急速に普及し，注目を集めている情報源として，Twitterを代表とするマイクロブログがある．Twitterでは，ユーザはツイートと呼ばれる140文字以内の短いメッセージを投稿する．また，ユーザはフォローと呼ばれる仕組みにより，興味を持ったユーザの最新のツイートをつねに受け取ることができる．この140文字以内という制限により，情報を気軽に発信できるため，ユーザは日々の出

[1] 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University, Nagoya, Aichi 464–8603, Japan
[a)] sasaki@cmplx.cse.nagoya-u.ac.jp

来事や趣味などの個人的な事柄に加えて，地震などの突発的な出来事に対する反応などをリアルタイムで投稿する．このように，ツイートにはユーザの属性や，突発的なイベントに関する情報が豊富に含まれており，これらの情報を抽出，活用することを目的とした研究は数多く行われている [6], [14], [15], [19].

Latent Dirichlet Allocation（LDA）[4] に代表されるトピックモデルは，文書中の潜在的意味を解析する手法として広く利用されている．また近年，Twitter に対して LDA を適用した研究も数多く報告され始めており，LDA を用いたユーザの分類 [14] や，影響力のあるユーザの推定 [19] などが報告されている．LDA は，文書が固有のトピック比率を持ち，それらのトピックを介して単語が生成されることを仮定した文書生成モデルである．LDA をツイート集合に適用する場合，一般的には 1 ツイートを 1 文書として適用する方法が考えられる．しかし，ツイートはニュース記事や新聞記事などの文書に比べると圧倒的に短く，またノイズとなる語句も多いため，直接 LDA を適用しても適切に機能しない．このため，従来研究では 1 ユーザの全ツイートを 1 文書として扱う方法が多く用いられている [9], [14], [19]．これに対し Zhao らは，Twitter の特徴を考慮し，1 ツイートが 1 トピックからなるという仮定に基づいた Twitter-LDA を提案し，前者の方法に比べて，トピックの意味のまとまりの面で優れていると報告している [21]．ただし Twitter-LDA では，LDA と同様サンプルの順序が交換可能であると仮定しているため，時間的な情報を考慮することができない．ユーザの興味は日々変化し，また Twitter 上の話題も実世界の動きとともに変化する．このような変化をとらえるためには，ツイートの投稿される時間的な順序を考慮する必要がある．また，これらのモデルは，新たなデータが得られるたびに全データを用いて学習を行う必要があるため，日々大量に蓄積されるツイート集合を効率的に解析することができない．これらの問題を解決するためには，時間発展を考慮し，かつオンライン学習を可能にすることが必要である．

本論文では，Twitter-LDA を改良し，ユーザごとに一般語とトピック語との割合を推定する新しいモデルを提案する．パープレキシティとコヒーレンスに基づく評価実験により，提案する改良モデルの妥当性を評価する．さらに，改良モデルに，オンラインストアにおけるユーザの購買行動をモデル化した Topic Tracking Model（TTM）[10] の機構を加えることで，ツイートの投稿される時間的な順序を考慮した，オンライン学習可能なトピックモデルである Twitter-TTM を提案する．提案モデルは，Twitter におけるユーザの興味と話題の時間発展を効率的にモデル化することが可能である．

## 2. Twitter-LDA の改良

### 2.1 改良モデル

Twitter-LDA のグラフィカルモデルを図 **1** に示す．Twitter-LDA は，1 つのツイートは 1 つのトピックに基づいて生成されるという仮定に基づいて，LDA を改良したモデルである．Twitter-LDA では，各ユーザがそれぞれ固有のトピック比率 $\phi_u$ を持つとする．ここで，$\phi_u$ はユーザ $u$ が各トピックに興味を持つ確率を表す．ユーザの各ツイートには，興味分布 $\phi_u$ に従って 1 つのトピックが割り当てられる．ツイートを構成する単語は，どのトピックにおいても出現するような単語の分布 $\theta_B$ から生成される一般語と，そのツイートの持つトピック $k$ の単語の分布 $\theta_k$ から生成されるトピック語からなる．各単語がどちらの分布から生成されるかは潜在変数 $y$ によって決まり，$y = 0$ ならば $\theta_B$ から，$y = 1$ ならば $\theta_k$ から生成される．潜在変数 $y$ は分布 $\pi$ から生成される．

Twitter-LDA では，潜在変数 $y$ の生成される分布 $\pi$ は，すべてのユーザに共通であり，どのユーザもトピック語と一般語を同じ割合で含むツイートを生成すると仮定しているが，これは実際のツイートの生成過程を適切には表現していないと考えられる．そこで本論文では，図 **2** に示すように，潜在変数 $y$ の分布 $\pi$ がユーザごとに異なるという仮定に基づいた改良モデルを提案する．改良モデルでは，ツイートを構成する一般語とトピック語の比率は，それぞれのユーザに固有の分布 $\pi_u$ から生成されると仮定する．これにより，改良モデルがより適切に実際のツイートの生成過程を表現できると考えられる．

### 2.2 評価実験

2.1 節で示した改良モデルの妥当性を評価するために，パープレキシティとコヒーレンススコアに基づき，LDA，Twitter-LDA，改良モデルの性能比較実験を行った．ただし，LDA は 1 ユーザの全ツイートを 1 文書として適用した．

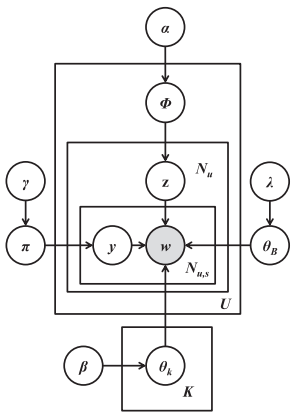


図 1　Twitter-LDA のグラフィカルモデル
**Fig. 1**　Graphical model of Twitter-LDA.

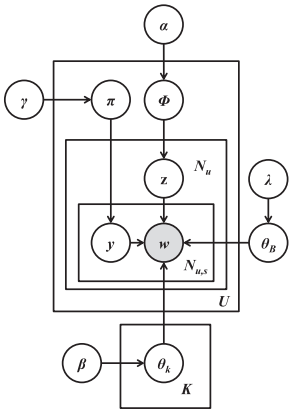


図 2　改良モデルのグラフィカルモデル
**Fig. 2**　Graphical model of improved-model.

実験には，2013 年 5 月 1 日に収集した，ユーザ数 16,411，ツイート数 493,462 の日本語ツイートデータを用いた．各ユーザは，2013 年 4 月 23 日の時点でツイート数が 1,000 以上のユーザからランダムに抽出した．各ツイートは，形態素解析を行った後，名詞だけを抽出し，データ全体での出現頻度が 30 回以上の単語のみとした．リツイート[*1]は，厳密にはそのユーザの発言とはいえないが，ユーザの興味を反映するものとして通常のツイートと同様に扱った．本前処理により，単語の存在しなくなったツイートを除いた結果，ユーザ数 16,170，ツイート数 371,199，語彙数 5,708 となった．以下の実験では，各モデルの学習には Collapsed ギブスサンプリング [7] を用い，反復回数は 500 回とした．また，各モデルのハイパーパラメータは，サンプリングが行われるたびに不動点反復法による尤度最大化によって推定した [13]．

#### 2.2.1　パープレキシティを用いた評価

パープレキシティは，学習によって得られたモデルが，テストデータ $D_{(test)}$ をどれだけ予測できるかを評価する指標である．

$$perplexity = \exp\Big(-\frac{1}{N}\sum_{u}\log p(\mathbf{w_u}|D_{(test)})\Big) \qquad (1)$$

ここで，$N$ はテストデータ中の全単語数であり，$\mathbf{w_u}$ はユーザ $u$ のツイートに含まれる全単語である．パープレキシティが低いほど，モデルの予測性能が高いことを示している．トピック数 $K$ を 50 から 250 まで 50 ずつ変化させ，それぞれ 10 分割交差検定を行った際のパープレキシティの結果を**表 1** に示す．なお表中の各値は，10 回試行におけるパープレキシティの平均値を，() 内の数値はその標準偏差をそれぞれ表している．

表 1 の結果から，改良モデルはどのトピック数においても Twitter-LDA に対してパープレキシティが改善されていることが分かる．このことから，一般語とトピック語の比率がユーザごとに異なるという仮定が適切であったと考えられる．なお LDA のパープレキシティがトピック数を増やしても下がらないのは，1 つのツイートが 1 つのトピックからなるという仮定がないためであると考えられる．

[*1] 他のユーザのツイートをそのまま引用してツイートすること．

表 1　各モデルにおけるパープレキシティ（10 回試行）
**Table 1**　Perplexity of each model in 10 runs.

| トピック数 | LDA | Twitter-LDA | 改良モデル |
|---|---|---|---|
| 50 | 943.5(9.8) | 1,120.4(24.4) | **725.9**(10.0) |
| 100 | 950.6(12.5) | 928.2(16.1) | **630.4**(12.5) |
| 150 | 958.1(6.2) | 825.3(10.8) | **580.3**(9.9) |
| 200 | 963.0(12.1) | 742.8(10.5) | **536.0**(9.5) |
| 250 | 973.4(9.9) | 695.4(11.9) | **507.4**(7.7) |

#### 2.2.2　コヒーレンスによる評価

それぞれのモデルにおいて生成されるトピックのコヒーレンス，すなわち意味的な一貫性を比較する．ここでは，文献 [12] において提案されているコヒーレンススコアに基づいて比較を行う．あるトピック $z$ における $p(v|z)$ の高い上位 $M$ 語 $V^{(z)} = (v_1^{(z)}, \ldots, v_M^{(z)})$ に対して，コヒーレンススコアは以下の式で定義される．

$$C(z; V^{(z)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{D(v_m^{(z)}, v_l^{(z)})+1}{D(v_l^{(z)})} \qquad (2)$$

式 (2) において，$D(v)$ は単語 $v$ が出現した文書の数であり，$D(v, v')$ は単語 $v$ と単語 $v'$ が共起した文書の数である．ただし，ここでは 1 ツイートを 1 文書としてこれらの値を算出する．

コヒーレンススコアは，ある概念（トピック）に属する単語は，同じ文書において互いに共起しやすいという考えに基づいている．この値が高いほど，トピックの意味的な一貫性が高いといえる．

各モデルにおけるトピックのコヒーレンスを評価するために，コヒーレンススコアの平均値

$$\frac{1}{K}\sum_{k}C(k; V^{(k)}) \qquad (3)$$

をモデルにおけるコヒーレンスとして用いる．トピック数 $K$ を 50 から 250 まで 50 ずつ変化させ，それぞれモデルの学習を 10 試行行ったときの式 (3) の値の平均値を**図 3** に示す．ただし，式 (2) における $M$ を 20 とした．

図 3 の結果から，LDA ではトピック数が増えるにつれてコヒーレンスが下がるのに対し，Twitter-LDA と改良モデルでは，コヒーレンスがほとんど下がらないことが分かる．これは，LDA においては，トピック数が増えると意味のないトピックが多く生成されることが原因であると考えられる．なおトピック数 $K = 50$ においては LDA のコヒーレンスが最も高いのは，上位の単語が出現頻度の高い一般語だけで構成されるトピックが多く生成されたためであると考えられる．一般語どうしはそもそも共起しやすいため，上位の単語が一般語だけであればコヒーレンススコアは高

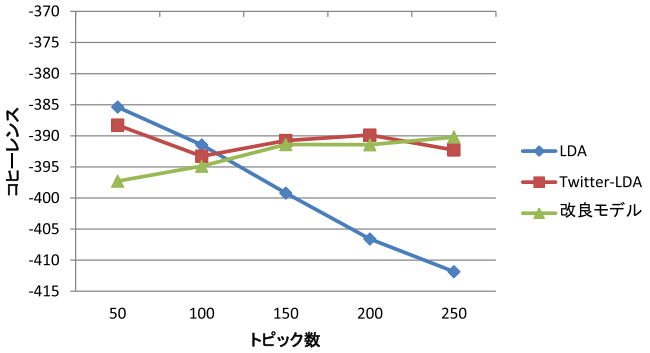


図 3　各モデルにおけるコヒーレンスの平均値（10 回試行）
**Fig. 3**　Coherence of each model averaged in 10 runs.

くなる．しかしトピック数を増やすと，上位の単語に一般語とトピック語が混在したトピックが多くなり，コヒーレンスが下がったと考えられる．これに対して Twitter-LDA と改良モデルでは，一般語とトピック語を区別しているため，各トピックの上位の単語がある概念に属するようなトピック語で構成されやすく，トピック数の増加はそれらの概念の粒度を小さくするのみであり，一般語とトピック語との混在にはつながらない．そのため，トピック数を増やしてもコヒーレンスが下がらなかったと考えられる．

## 3. 提案モデル

### 3.1 Topic Tracking Model に基づくモデル拡張

2.1 節で示した改良モデルを，オンライン学習の可能なトピックモデルに拡張する．本論文では，時間発展を考慮したトピックモデルである，Topic Tracking Model（TTM）[10] の機構に基づいてモデルの拡張を行う．

TTM は，ユーザの購買行動における興味とトピックの時間発展を追跡することができるトピックモデルである．時間発展を考慮したトピックモデルとしては，TTM のほかに Dynamic Topic Model（DTM）[3] や，Topic over Time（ToT）[17] などがある．DTM は時系列文書集合におけるトピックの時間発展を解析するためのモデルである．DTM のグラフィカルモデルを**図 4** に示す．図 4 のように，DTM では，個々のユーザの興味の時間発展を考慮する機構がない．また，ToT はオンライン学習ができないため，日々大量に生成されるツイート集合を効率良くモデル化することができない．本論文では，ユーザの興味とトピックの両方の時間発展を考慮し，かつオンライン学習が可能であることが，ツイート集合のモデル化に必要であると考え，TTM の機構を改良モデルに導入する．**図 5** に，TTM のグラフィカルモデルを示す．

TTM では，ユーザの興味分布の平均は，新たなデータが観測されない場合，1 時刻前の興味分布と同じであると仮定する．時刻 $t$ におけるユーザ $u$ の興味分布 $\phi_{t,u}$ は，平均が 1 時刻前の興味分布の推定値 $\hat{\phi}_{t-1,u}$ であり，精度（分散の逆数）が $\alpha_{t,u}$ である以下のディリクレ事前分布に従って生成される．

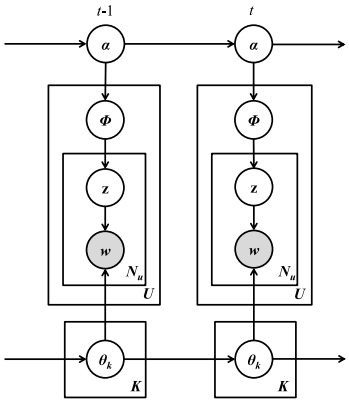


図 4　DTM のグラフィカルモデル
**Fig. 4**　Graphical model of DTM.

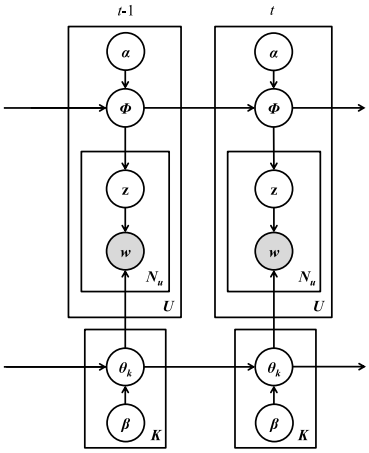


図 5　TTM のグラフィカルモデル
**Fig. 5**　Graphical model of TTM.

$$p(\phi_{t,u}|\hat{\phi}_{t-1,u},\alpha_{t,u}) \propto \prod_k \phi_{t,u,k}^{\alpha_{t,u}\hat{\phi}_{t-1,u,k}-1} \tag{4}$$

式 (4) において，$\phi_{t,u,k}$ は $\phi_{t,u}$ におけるトピック $k$ への興味確率であり，精度 $\alpha_{t,u}$ は，ユーザ $u$ の時刻 $t-1$ と $t$ の間での興味の一貫性を表している．$\alpha_{t,u}$ が大きいほど，興味の分散が小さく，興味の一貫性が高いことを意味する．時刻 $t$ におけるトピック $k$ の単語分布 $\theta_{t,k}$ についても同様に，平均が 1 時刻前の単語分布の推定値 $\hat{\theta}_{t-1,k}$，精度が $\beta_{t,u}$ である以下のディリクレ事前分布から生成される．なお式 (5) において，$\theta_{t,k,v}$ は $\theta_{t,k}$ における単語 $v$ の生成確率である．

$$p(\theta_{t,k}|\hat{\theta}_{t-1,k},\beta_{t,k}) \propto \prod_v \theta_{t,k,v}^{\beta_{t,k}\hat{\theta}_{t-1,k,v}-1} \tag{5}$$

本論文では，上記の TTM の機構を，2.1 節で示した改良モデルに取り入れたモデル（Twitter-TTM）を提案する．すなわち，TTM と同様に，ユーザの興味分布 $\phi_{t,u}$ とトピックの単語分布 $\theta_{t,k}$ が，1 時刻前に推定した分布に依存するように改良モデルの拡張を行う．ただし，一般語の分布 $\theta_B$ と，ユーザにおける一般語とトピック語の比率を

(1) Draw $\theta_{t,B}$ ∼Dirichlet($\lambda$)
(2) For each topic $k = 1, ..., K$,
　(a) draw $\theta_{t,k}$ ∼Dirichlet($\beta_{t,k}\hat{\theta}_{t-1,k}$)
(3) For each user $u = 1, ..., U$,
　(a) draw $\phi_{t,u}$ ∼Dirichlet($\alpha_{t,u}\hat{\phi}_{t-1,u}$)
　(b) draw $\pi_{t,u}$ ∼Beta($\gamma$)
　(c) for each tweet $s = 1, ..., N_u$
　　(i) draw $z_{t,u,s}$ ∼Multinomial($\phi_{t,u}$)
　　(ii) for each word $n = 1, ..., N_{u,s}$
　　(A) draw $y_{t,u,s,n}$ ∼Bernoulli($\pi_{t,u}$)
　　(B) draw $w_{t,u,s,n}$ ∼
　　Multinomial($\theta_{t,B}$) if $y_{t,u,s,n} = 0$ or
　　Multinomial($\theta_{t,z_{t,u,s}}$) if $y_{t,u,s,n} = 1$

図 6　提案モデルにおけるツイートの生成過程
**Fig. 6**　Generation process of tweets in proposed model.

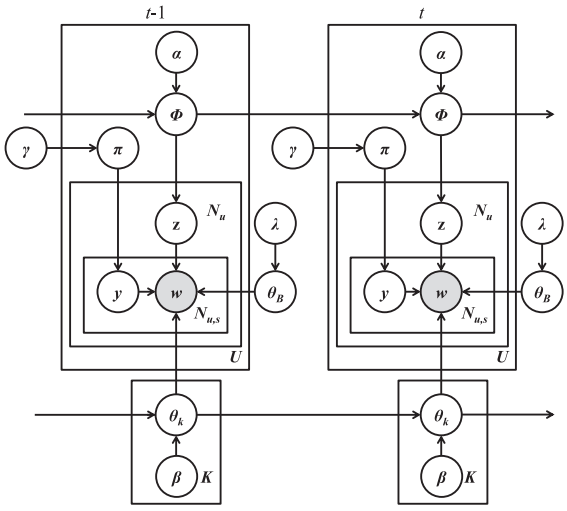


図 7　Twitter-TTM のグラフィカルモデル
**Fig. 7**　Graphical model of Twitter-TTM.

与える分布 $\pi_u$ については，1 時刻前への依存性は考慮しない．これは，これらの分布については時間的な依存性がない，もしくはきわめて小さいと考えられるためである．

提案モデルの生成過程とグラフィカルモデルをそれぞれ**図 6**，**図 7** に示す．提案モデルにより，Twitter の特徴を考慮したうえで，ユーザの興味およびトピックの時間発展を逐次推定することが可能となる．なお，提案モデルは文献 [10] と同様に，1 時刻前の依存性だけでなく，長期的な依存性も考慮できるように拡張することが可能である．

### 3.2　モデルの学習

提案モデルの学習には，Collapsed ギブスサンプリング [7] による潜在変数の推定と，完全尤度の最大化によるパラメータの推定を交互に繰り返す確率的 EM アルゴリズム [16] を用いる．時刻 $t$ における興味分布の集合 $\mathbf{\Phi}_t = \{\hat{\phi}_{t,u}\}_{u=1}^{U}$，トピックの単語分布の集合 $\mathbf{\Theta}_t = \{\hat{\theta}_{t,k}\}_{k=1}^{K}$，一般語分布 $\theta_{t,B}$，一般語とトピック語の比率を与える分布 $\pi_{t,u}$，興味精度 $\boldsymbol{\alpha_t} = \{\alpha_{t,u}\}_{u=1}^{U}$，流行精度 $\boldsymbol{\beta_t} = \{\beta_{t,k}\}_{k=1}^{K}$ を，時刻 $t$ において観測されたデータと 1 時刻前の興味分布集合 $\hat{\mathbf{\Phi}}_{t-1}$，単語分布集合 $\hat{\mathbf{\Theta}}_{t-1}$ から推定する．

ギブスサンプリングにおける潜在変数のサンプリング確率は，データと潜在変数の同時確率（完全尤度）から導出される．時刻 $t$ におけるツイートデータ，潜在トピック $z$ の集合，潜在変数 $y$ の集合をそれぞれ $D_t$，$Z_t$，$Y_t$ とすると，完全尤度は以下の式で表される．

$$
\begin{aligned}
&p(D_t, Y_t, Z_t|\hat{\mathbf{\Phi}}_{t-1}, \hat{\mathbf{\Theta}}_{t-1}, \boldsymbol{\alpha_t}, \boldsymbol{\beta_t}, \lambda, \gamma) \\
&= \left(\frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2}\right)^U \prod_u \frac{\Gamma(\gamma + n_{t,u,B})\Gamma(\gamma + n_{t,u,K})}{\Gamma(2\gamma + n_{t,u})} \\
&\times \frac{\Gamma(V\lambda)}{\Gamma(\lambda)^V} \frac{\prod_v \Gamma(n_{t,B,v} + \lambda)}{\Gamma(n_{t,B} + V\lambda)} \\
&\times \prod_k \frac{\Gamma(\beta_{t,k})}{\Gamma(n_{t,k} + \beta_{t,k})} \prod_v \frac{\Gamma(n_{t,k,v} + \beta_{t,k}\hat{\theta}_{t-1,k,v})}{\Gamma(\beta_{t,k}\hat{\theta}_{t-1,k,v})} \\
&\times \prod_u \frac{\Gamma(\alpha_{t,u})}{\Gamma(c_{t,u} + \alpha_{t,u})} \prod_k \frac{\Gamma(c_{t,u,k} + \alpha_{t,u}\hat{\phi}_{t-1,u,k})}{\Gamma(\alpha_{t,u}\hat{\phi}_{t-1,u,k})} \qquad (6)
\end{aligned}
$$

ここで，$n_{t,u,B}$，$n_{t,u,K}$ はそれぞれ，時刻 $t$ におけるユーザ $u$ の一般語，トピック語の数，$n_{t,B,v}$ は単語 $v$ が時刻 $t$ で一般語に割り当てられた回数，$n_{t,k,v}$ は単語 $v$ が時刻 $t$ でトピック $k$ に割り当てられた回数，$c_{t,u,k}$ は時刻 $t$ におけるユーザ $u$ のツイートのうち，トピック $k$ に割り当てられたものの数である．また，$n_{t,u} = n_{t,u,B} + n_{t,u,K}$，$n_{t,B} = \sum_v n_{t,B,v}$，$n_{t,K} = \sum_k n_{t,k} = \sum_k \sum_v n_{t,k,v}$，$c_{t,u} = \sum_k c_{t,u,k}$ である．

式 (6) を用いて，潜在トピック $z$ のサンプリング確率は以下のように導出することができる．

$$
\begin{aligned}
&p(z_i = k|D_t, Y_t, Z_{t\backslash i}, \hat{\mathbf{\Phi}}_{t-1}, \hat{\mathbf{\Theta}}_{t-1}, \boldsymbol{\alpha_t}, \boldsymbol{\beta_t}) \\
&\propto \frac{c_{t,u,k\backslash i} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}}{c_{t,u\backslash i} + \alpha_{t,u}} \\
&\times \frac{\Gamma(n_{t,k\backslash i} + \beta_{t,k})}{\Gamma(n_{t,k} + \beta_{t,k})} \prod_v \frac{\Gamma(n_{t,k,v} + \beta_{t,k}\hat{\theta}_{t-1,k,v})}{\Gamma(n_{t,k,v\backslash i} + \beta_{t,k}\hat{\theta}_{t-1,k,v})} \qquad (7)
\end{aligned}
$$

ここで，$i = (t, u, s)$ であり，$z_i$ は時刻 $t$ におけるユーザ $u$ のツイート $s$ に割り当てられるトピックである．また，$\backslash i$ はツイート $i$ を除いたときの数であることを表す．

また，$z_i = k$ のとき，潜在変数 $y$ のサンプリング確率は以下の式で与えられる．

$$
\begin{aligned}
&p(y_j = 0|D_t, Y_{t\backslash j}, Z_t, \lambda, \gamma) \\
&\propto \frac{n_{t,B,v\backslash j} + \lambda}{n_{t,B\backslash j} + V\lambda} \frac{n_{t,u,B\backslash j} + \gamma}{n_{t,u\backslash j} + 2\gamma} \qquad (8) \\
&p(y_j = 1|D_t, Y_{t\backslash j}, Z_t, \hat{\mathbf{\Theta}}_{t-1}, \boldsymbol{\beta_t}, \gamma) \\
&\propto \frac{n_{t,k,v\backslash j} + \beta_{t,k}\hat{\theta}_{t-1,k,v}}{n_{t,k\backslash j} + \beta_{t,k}} \frac{n_{t,u,K\backslash j} + \gamma}{n_{t,u\backslash j} + 2\gamma} \qquad (9)
\end{aligned}
$$

ここで，$j = (t, u, s, n)$ であり，$y_j$ は時刻 $t$ におけるユーザ $u$ の $s$ 番目のツイート中の $n$ 番目の単語に割り当てられる潜在変数である．また，$\backslash j$ は単語 $j$ を除いたときの数であることを表す．

精度 $\boldsymbol{\alpha_t}$，$\boldsymbol{\beta_t}$ は，不動点反復法による完全尤度の最大化によって推定する [13]．このとき，$\alpha_{t,u}$ の更新式は

$$\alpha_{t,u}^{new} = \alpha_{t,u} \times \frac{\sum_k \hat{\phi}_{t-1,u,k}(\Psi(c_{t,u,k}+\alpha_{t,u}\hat{\phi}_{t-1,u,k})-\Psi(\alpha_{t,u}\hat{\phi}_{t-1,u,k}))}{\Psi(c_{t,u}+\alpha_{t,u})-\Psi(\alpha_{t,u})} \tag{10}$$

また，$\beta_{t,k}$ の更新式は

$$\beta_{t,k}^{new} = \beta_{t,k} \times \frac{\sum_v \hat{\theta}_{t-1,k,v}(\Psi(n_{t,k,v}+\beta_{t,k}\hat{\theta}_{t-1,k,v})-\Psi(\beta_{t,k}\hat{\theta}_{t-1,k,v}))}{\Psi(n_{t,k}+\beta_{t,k})-\Psi(\beta_{t,k})} \tag{11}$$

となる．

上記のギブスサンプリングによる潜在トピックおよび潜在変数の推定と，完全尤度最大化によるパラメータの推定を繰り返すことにより，提案モデルの学習が行われる．十分回数反復を行った後，$\phi_{t,u,k}$ と $\theta_{t,k,v}$ の推定値は，MAP 推定により以下の式で求めることができる．

$$\hat{\phi}_{t,u,k} = \frac{c_{t,u,k}+\alpha_{t,u}\hat{\phi}_{t-1,u,k}}{c_{t,u}+\alpha_{t,u}}, \tag{12}$$

$$\hat{\theta}_{t,k,v} = \frac{n_{t,k,v}+\beta_{t,k}\hat{\theta}_{t-1,k,v}}{n_{t,k}+\beta_{t,k}} \tag{13}$$

また，これらの推定値は次の時刻 $t+1$ におけるモデル学習に用いられる．

## 4. 関連研究

近年，Twitter を対象とした様々なトピックモデルが提案されている．Diao らは，Twitter-LDA に，ユーザが個人的な興味と時間に依存したイベントのどちらかについてツイートをするという仮定を加えたトピックモデルを提案し，過去に起きた突発的なイベントに関するトピックの抽出を行っている [6]．Yan らは，ツイートのような短いテキストを対象として，各トピックから 2 つの単語の対が生成されるという仮定に基づいた biterm topic model（BTM）を提案している [20]．Chua らは，イベントの要約のためのツイート抽出を目的として，限られた時間窓内において，ツイートの生成された順序と時間間隔を考慮した 2 つのモデルを提案している [5]．しかしこれらのモデルはいずれも，時間発展を考慮しておらず，またオンライン学習ができないため，日々大量に生成されるツイートを効率的にモデル化することができない．提案モデルでは，ユーザの興味とトピックの時間発展を考慮し，オンライン学習が可能である点でこれらのモデルと異なる．

オンライン学習が可能なモデルとしては，文献 [18] において Wang らにより提案されている TM-LDA がある．TM-LDA では，ユーザのツイートのトピック分布が，次元数がトピック数に等しい正方行列を介して次の時刻のトピック分布へと変遷すると仮定し，ユーザの興味の変化をオンラインで学習可能としている．しかし TM-LDA では，トピックの時間発展については考慮していない．また Lau らは，オンライン学習が可能である Online LDA（OLDA）[2] を，語彙数の変化も考慮できるよう拡張し，Twitter におけるトレンドの解析に適用している [11]．また，LDA を変分ベイズ法を用いてオンラインで学習する方法も提案されている [8]．ただしこれらの研究では，いずれも LDA をベースにしているため，ツイートの短さに対する工夫がモデルに組み込まれていない．これに対し提案モデルでは，1 ツイートを 1 トピックと仮定してモデルを構築することで，ツイートの短さに対応している．この仮定は，ツイートは一般的な文書に比べて非常に短いため，1 つのツイートに複数のトピックが存在することは少なく，たいてい 1 つのトピックについて書かれているという仮説に基づいている．

## 5. 評価実験

提案モデルの有効性を評価するため，実際のツイートデータを用いた評価実験を行った．実験には，2013 年 5 月 1 日から 5 月 15 日までの間に収集した，ユーザ数 19,764（2013 年 4 月 23 日の時点でツイート数が 1,000 以上），全ツイート数 10,244,572 の日本語ツイートデータを用いた．各ツイートは，リツイートも含めて形態素解析を行った後，名詞だけを抽出し，データ全体での出現頻度が 30 回以上の単語のみとした．本前処理により，ユーザ数 19,729，全ツイート数 8,257,368，語彙数 63,030 となった．

提案モデルの評価として，LDA，TTM，改良モデルとの予測性能の比較を行った．トピック数 $K$ は 100 とし，3.2 節で示したモデル学習における反復回数は 500 とした．各モデルのハイパーパラメータは，尤度最大化により逐次推定した [13]．また，時刻の単位は 1 日とし，評価指標としては 2.2 節でも用いたパープレキシティを用いた．LDA と改良モデルは，それぞれの時刻のツイートデータのみを用いて学習を行った．

**図 8** に 5 月 2 日から 5 月 15 日までの，各時刻のパープレキシティの結果を示す．ここで，$x$ 軸は $t=1$ が 5 月 2 日に，$t=2$ が 5 月 3 日に，... $t=14$ が 5 月 15 日にそれ

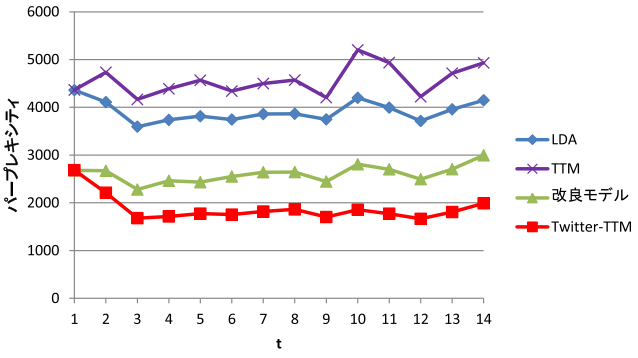


**図 8**　各時刻におけるパープレキシティ
**Fig. 8**　Perplexity for each time.

ぞれ対応している．各時刻におけるパープレキシティは，時刻 $t-1$ において学習したモデルの，時刻 $t$ におけるデータに対する値であり，次の時刻のツイートに対する予測性能を示している．なお $t=1$ においては，TTM と提案モデルの学習は，それぞれ LDA と改良モデルの学習と同等である．

図 8 より，提案モデルがすべての時刻において TTM よりもパープレキシティが低い，すなわち予測性能が高いことが分かる．また提案モデルは，$t=2$ 以降，改良モデルよりもパープレキシティが改善されていることが分かる．これは，時間発展を考慮することにより，適切に学習が行われたためであると考えられる．一方 TTM は，時間発展を考慮しているにもかかわらず，時間発展を考慮しない通常の LDA よりもパープレキシティが低くなっていることが分かる．これは，LDA による学習がそもそも適切に行われていないために，1 時刻前の推定値がむしろノイズとなってしまい，適切な学習が行われなかったためであると考えられる．

以上の実験結果から，提案モデルは従来法に比べて予測性能が高く，日々蓄積されるツイートをより適切にモデル化できることを確認した．

## 6. おわりに

本論文では，Twitter-LDA を改良し，ユーザごとに一般語とトピック語との割合を推定する改良モデルを提案した．さらに，ユーザの購買行動をモデル化したトピックモデルである TTM の機構を改良モデルに加えることにより，Twitter におけるユーザの興味と話題の時間発展を考慮した，オンライン学習が可能なトピックモデル Twitter-TTM を提案した．日本語ツイートデータを用いた評価実験により，提案モデルは従来モデルよりもツイートの予測性能が高く，より適切にツイート集合をモデル化できることを確認した．

今後の課題としては，時間の推移によるトピックの発生や消滅を考慮する [1] ことがあげられる．また，Twitter におけるリアルタイムでのトレンド解析や，ユーザへのコンテンツ推薦などに適用し，提案モデルの有効性に対するさらなる検証を行っていくことなどがあげられる．

### 参考文献


[1] Ahmed, A. and Xing, E.P.: Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream, *Proc. UAI '10* (2010).

[2] AlSumait, L., Barbara, D. and Domeniconi, C.: Online LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking, *Proc. ICDM '08*, pp.3–12, IEEE (2008).

[3] Blei, D.M. and Lafferty, J.D.: Dynamic topic models, *Proc. ICML '06*, pp.113–120, ACM (2006).

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

[5] Chua, F. and Asur, S.: Automatic Summarization of Events from Social Media, Technical report, HP Labs.

[6] Diao, Q., Jiang, J., Zhu, F. and Lim, E.: Finding bursty topics from microblogs, *Proc. ACL '12*, pp.536–544, ACL (2012).

[7] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. National Academy of Sciences of the United States of America*, Vol.101, No.Suppl 1, pp.5228–5235 (2004).

[8] Hoffman, M., Bach, F.R. and Blei, D.M.: Online learning for latent Dirichlet allocation, *Proc. NIPS '10*, pp.856–864 (2010).

[9] Hong, L. and Davison, B.D.: Empirical study of topic modeling in twitter, *Proc. SOMA '10*, pp.80–88, ACM (2010).

[10] Iwata, T., Watanabe, S., Yamada, T. and Ueda, N.: Topic Tracking Model for Analyzing Consumer Purchase Behavior, *Proc. IJCAI '09*, Vol.9, pp.1427–1432 (2009).

[11] Lau, J.H., Collier, N. and Baldwin, T.: On-line Trend Analysis with Topic Models:# twitter Trends Detection Topic Model Online, *Proc. COLING '12*, pp.1519–1534 (2012).

[12] Mimno, D., Wallach, H.M., Talley, E., Leenders, M. and McCallum, A.: Optimizing semantic coherence in topic models, *Proc. EMNLP '11*, pp.262–272, ACL (2011).

[13] Minka, T.: Estimating a Dirichlet distribution, Technical report, MIT (2000).

[14] Pennacchiotti, M. and Popescu, A.M.: A Machine Learning Approach to Twitter User Classification, *Proc. ICWSM '11*, pp.281–288 (2011).

[15] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors, *Proc. WWW '10*, pp.851–860, ACM (2010).

[16] Wallach, H.M.: Topic modeling: Beyond bag-of-words, *Proc. ICML '06*, pp.977–984, ACM (2006).

[17] Wang, X. and McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends, *Proc. KDD '06*, pp.424–433, ACM (2006).

[18] Wang, Y., Agichtein, E. and Benzi, M.: TM-LDA: Efficient Online Modeling of the Latent Topic Transitions in Social Media, *Proc. KDD '12*, pp.123–131, ACM (2012).

[19] Weng, J., Lim, E.P., Jiang, J. and He, Q.: Twitterrank: Finding topic-sensitive influential twitterers, *Proc. WSDM '10*, pp.261–270, ACM (2010).

[20] Yan, X., Guo, J., Lan, Y. and Cheng, X.: A Biterm Topic Model for Short Texts, *Proc. WWW '13*, pp.1445–1456, International World Wide Web Conferences Steering Committee (2013).

[21] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H. and Li, X.: Comparing twitter and traditional media using topic models, *Proc. Advances in Information Retrieval*, pp.338–349, Springer (2011).

**佐々木 謙太朗**

2013 年 3 月名古屋大学工学部電気電子・情報工学科卒業．同年 4 月同大学大学院工学研究科博士課程前期課程計算理工学専攻に入学，現在に至る．主として自然言語処理に関する研究に従事．

**吉川 大弘**　（正会員）

1997 年名古屋大学大学院博士課程修了．同年カリフォルニア大学バークレー校ソフトコンピューティング研究所客員研究員．1998 年三重大学工学部助手．2005 年名古屋大学大学院工学研究科 COE 特任准教授．2006 年 10 月同研究科准教授，現在に至る．主としてソフトコンピューティングとその応用に関する研究に従事．博士（工学）．IEEE，電子情報通信学会，日本知能情報ファジィ学会，進化計算学会各会員．

**古橋 武**

1985 年名古屋大学大学院工学研究科博士後期課程電気系専攻修了．工学博士．2004 年名古屋大学大学院工学研究科計算理工学専攻教授，現在に至る．ソフトコンピューティング，感性工学に関する研究に従事．1996 年日本ファジィ学会論文賞受賞．IEEE，日本知能情報ファジィ学会，電気学会等の各会員．