D-031

Twitter に適したオンライン学習可能なトピックモデルの検討
A Study on Online Topic Model for Twitter

佐々木　謙太朗 †　　吉川 大弘 †　　古橋 武 †
Kentaro Sasaki　Tomohiro Yoshikawa　Takeshi Furuhashi

1. はじめに

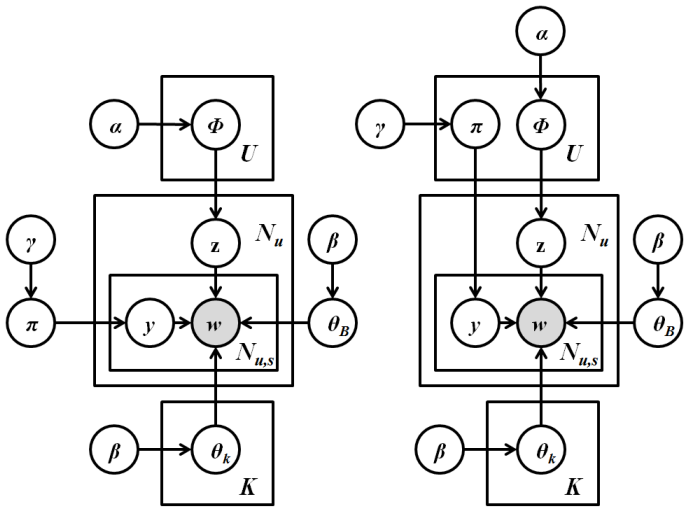
近年急速に普及し，注目を集めている情報源として，Twitter を代表とするマイクロブログがある．一方，Latent Dirichlet Allocation (LDA)[1] は，様々な分野で応用されているトピックモデルであり，Twitter に対して適用した研究も数多く報告され始めている [2][3]．Twitter では，ユーザはツイートと呼ばれる 140 文字以内の短いメッセージを投稿する．この 140 文字以内という制限により，ユーザが気軽に情報を発信できるため，Twitter は速報性，リアルタイム性が極めて高い情報源である．しかしその反面，ツイートの短さと，ノイズの多さのため，従来の自然言語処理技術が有効に働かないという問題点が存在する．そのため，上述の LDA を直接適用しても適切には機能せず，通常 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，あるいは 1 ユーザの全ツイートを 1 文書とする方法がとられる．これに対して，Twitter の特徴を考慮し，1 ツイートが 1 トピックから成るという仮定に基づいた Twitter-LDA が提案されている [4]．ただし，このモデルでは，通常の LDA と同様に時間発展を考慮しておらず，またデータが更新されるごとに全データを用いて再度学習を行う必要があるため，オンライン学習も困難である．そのため，Twitter 上のトピックの時間発展を追跡することはできない．

本稿ではまず，Twitter-LDA を改良し，ユーザごとにトピック語と一般語との割合を推定する新しいモデルを考案する．パープレキシティを用いた実験により，その改良モデルの妥当性を評価する．さらに，考案したモデルを拡張し，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．

2. Twitter-LDA の改良

図 1(a) に，Twitter-LDA のグラフィカルモデルを示す．Twitter-LDA では，一つの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一般語の分布 θ_B と，そのツイートの持つトピック k の単語分布 θ_k のどちらか一方から生成されることを仮定している．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 θ_B から，$y=1$ ならば θ_k から生成される．

Zhao らは，[4] において LDA との比較の際，Twitter-LDA は各トピックにおける出現確率の高い語にノイズ（一般語）が少ないことを実験により示している．しかし我々の行った実験では，Twitter-LDA は，言語モデルの性能に対する評価尺度として一般的に用いられるパープレキシティ[1] の上では，LDA よりも悪くなるという結果になった．これは，潜在変数 y の生成される分布 π が，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するという仮定を置いているためだと考えられる．この仮定は実際のツイートの生成過程をうまく表現しておらず，ユーザによってトピック語と一般語の割合は変化すると仮定する方がより適切であると考えられる．そこで本稿では，図 1(b) に示すように，潜在変数 y の分布 π がユーザごとに異なるという仮定に基づいたモデルを提案する．



(a)Twitter-LDA　　(b) 改良モデル

図 1: グラフィカルモデル

3. パープレキシティ評価実験

2 節で示した Twitter-LDA の改良モデルの妥当性を評価するために，パープレキシティに基づく評価実験を行った．パープレキシティは，学習によって得られたモデルが，実際に観測された単語をどれだけ予測出来るかを評価する指標である．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．実験には, 2013 年 3 月 13 日に収集したユーザ数 409, ツイート数 9176, 語彙数 2660 の日本語ツイートデータを用いた．このツイートデータのうち，90% のツイートを学習に用い，残り 10% をパープレキシティの算出に用いた．各モデルのハイパーパラメータはすべて 0.1 とした．また，学習にはギブスサンプリングを用い，反復回数は 500 とした．なお，LDA は各ユーザの全ツイートを 1 文書として学習を行った．LDA，Twitter-LDA，提案モデルに対して，トピック数を 10 〜100 まで 10 ずつ変化させ，それぞれ 10 試行学習を行った際のパープレキシティの平均値を図 2 に示す．

結果から，提案する改良モデルは，パープレキシティにおいて LDA と Twitter-LDA の両方に対して改善していることがわかる．このことから，提案するモデルが Twitter の特徴をより適切に反映しており，一般語とトピック語の割合がユーザごとに異なるという仮定が妥当であると考えられる．

4. 提案モデル

2 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．本稿では，時間発展を考慮し，オンライン学習可能なトピックモデルの一つである，Topic Tracking Model (TTM)[5] の機構に基づいてモデルの拡張を行う．

TTM は，時間変化するユーザの興味と，トピックの発展を追跡することができるトピックモデルである．TTM において，時刻 t におけるユーザ u の興味分布 $\phi_{t,u}$ は，平均が一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

†名古屋大学大学院工学研究科計算理工学専攻

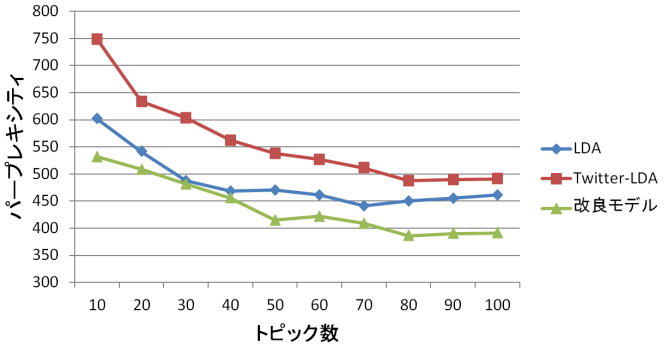


図 2: パープレキシティの比較

$\alpha_{t,u}$ である以下のディリクレ事前分布に従って生成される．

$$P(\phi_{t,u}|\hat{\phi}_{t-1,u},\alpha_{t,u}) \propto \prod_{k} \phi_{t,u,k}^{\alpha_{t,u}\hat{\phi}_{t-1,u,k}-1} \quad (1)$$

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，平均が一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．

$$P(\theta_{t,k}|\hat{\theta}_{t-1,k},\beta_{t,k}) \propto \prod_{v} \theta_{t,k,v}^{\beta_{t,k}\hat{\theta}_{t-1,k,v}-1} \quad (2)$$

本稿では，以上に述べた TTM の機構を，2 節で述べた改良モデルに取り入れたモデルを提案する．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 3，図 4 に示す．提案モデルにより，Twitter の特徴を考慮した上でユーザの興味およびトピックの流行の変化を逐次推定することが可能となる．

5. おわりに

本稿では，Twitter に適したオンライン学習が可能なトピックモデルを提案した．初めに，Twitter 向けに LDA を改良したトピックモデルである Twitter-LDA に対して，一般語とトピック語の割合はユーザごとに異なるという仮定を新たに加えた改良モデルを考案した．実験の結果から，改良モデルは Twitter-LDA や LDA と比較してパープレキシティが改善されることを示した．本稿ではさらに，この改良モデルに対して，時間発展を考慮したトピックモデルである TTM の機構を取り入れた新しいモデルを提案した．

今後は，提案モデルを Twitter におけるリアルタイムでの話題の検出，時間発展の追跡などに適用し，既存のモデルとの比較評価を行っていく予定である．

1. Draw $\theta_{t,B}$ ∼Dirichlet(λ)
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(γ)
   (c) for each tweet $s = 1, ..., N_u$
      i. draw $z_{t,u,s}$ ∼Multinomial($\phi_{t,u}$)
      ii. for each word $v = 1, ..., N_{u,s}$
         A. draw $y_{t,u,s,v}$ ∼Bernoulli($\pi_{t,u}$)
         B. draw $w_{t,u,s,v}$ ∼ Multinomial($\theta_{t,B}$) if $y_{t,u,s,v} = 0$ or Multinomial($\theta_{t,z_{t,u,s}}$) if $y_{t,u,s,v} = 1$

図 3: 提案モデルにおけるツイートの生成過程

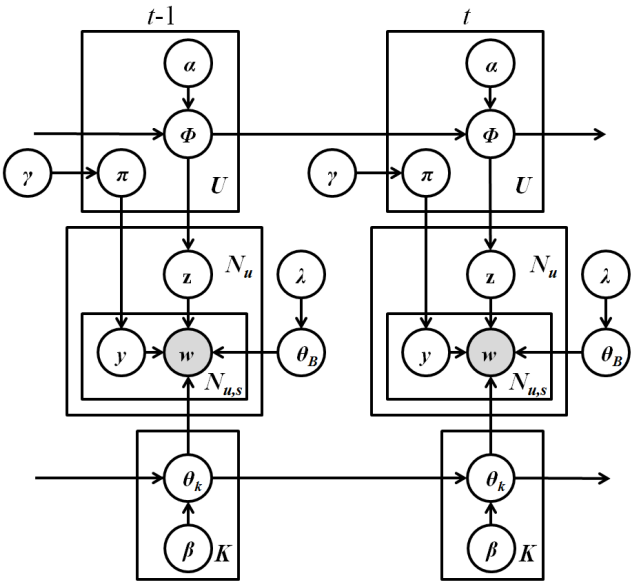


図 4: 提案モデル

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan: Latent dirichlet allocation, Machine Learning Research, Vol. 3, pp. 993-1022, 2003

[2] Marco Pennacchiotti and Siva Gurumurthy: Investigating topic models for social media user recommendation, In WWW2011, pp. 101-102, 2011

[3] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He: Twitterrank: finding topic-sensitive influential twitterers, In WSDM 2010, 2010

[4] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li: Comparing twitter and traditional media using topic models, In Proceedings of the 33rd European conference on Advances in information retrieval, pp. 338-349, 2011

[5] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda: Topic tracking model for analyzing consumer purchase behavior, in Proc. IJCAI, pp. 1427-1432, 2009