2F3-4

マイクロブログ解析のための混合トピックモデル

Mixture of Topic Models for Analysing Microblogs

今井 優作 *1 Yusaku Imai　岩田 具治 *2 Tomoharu Iwata　澤田 宏 *3 Hiroshi Sawada　山田 武士 *2 Takeshi Yamada

*1奈良先端科学技術大学院大学 Nara Institute of Science and Technology　*2NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

*3NTT サービスエボリューション研究所 NTT Service Evolution Laboratories

Topic Models are widely used for analysing large-scale text information. In some studies, for analysing microblogs such as Twitter, all the tweets of each user are aggregated as a single document, because tweets are too short and can not analyse them properly. As the result, the number of words is increased, but the difference of topics can not be expressed properly. In this paper, we propose a new topic model to overcome these difficulties. The proposed model clusters a set of tweets for each user. The tweets assigned to a same cluster are considered as a single document, and we infer topic proportions for each cluster. Because the proposed method has a topic distribution for each cluster, we can express a tweet as a mixture of topic distributions. In the experiment, we demonstrate the effectiveness of the proposed model using dataset of Twitter.

1. はじめに

近年，Twitter を代表とするマイクロブログが急速に普及し，ビジネスや研究分野において注目を浴びている．現在，全世界で 2 億人以上の人々が Twitter に登録し，ユーザは 140 字以内の「ツイート」と呼ばれる短文を投稿することで日常の出来事や趣味などの個人的な事柄を他人と共有できる．

大規模なテキスト情報から知識を獲得するための統計的モデリング手法としてトピックモデル [Hofmann 99] が広く利用されており，Twitter に対して適用した研究も多く報告されている．Weng らは潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation)[Blei 03] を用いて影響力のあるユーザを推定する方法を提案している [Weng 10]．また，Pennacchiotti らはツイート情報による LDA を用いたユーザの分類モデルを提案している [Pennacchiotti 11]．これらの先行研究では，ツイートが非常に短文であるために適切にモデル化できないことから，1 ツイートを 1 文書とするのではなく，各ユーザの全ツイートを擬似的に 1 文書として扱う方法を用いている．この方法により 1 文書に含まれる単語数を多くできるが，文書毎のトピックの違いを表現できないという問題がある．この問題に対し，Zhao らは Twitter の特徴を考慮し，1 ツイートが 1 トピックから成るという仮説を元に Twitter-LDA を提案している [Zhao 11]．Twitter-LDA は，ツイートの短さによって適切にモデル化できない問題を解消し，よりまとまりのあるトピックを抽出できる．しかし，Twitter-LDA により文書毎のトピックの違いを表現できるが，1 ツイートが複数のトピックから成るようなデータを表現できない．本稿では，各ユーザのツイート集合を複数のクラスタに分割し，同じクラスタに割り当てられたツイート集合を 1 文書とみなすことで，クラスタ毎に 1 つのトピック分布をもつトピックモデルを提案する．提案モデルにより，1 文書に含まれる単語数が短い問題，および文書毎のトピックを表現できない問題を解決し，かつ複数のトピックから成るツイートもモデル化できる．実験により，提案モデルが高い精度でツイート集合をモデル化できることを示す．

連絡先: 今井優作，奈良先端科学技術大学院大学情報科学研究科，imai.yusaku.is7@is.naist.jp

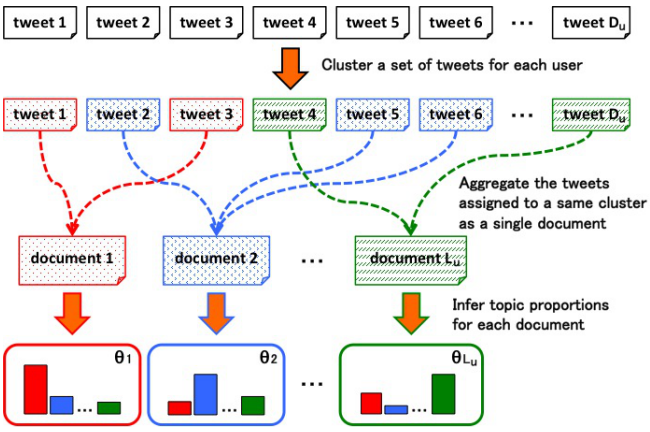

図 1: 提案法の概要図

2. 提案法

2.1 混合トピックモデル

本稿では，マイクロブログ解析のためのトピックモデルとして，混合トピックモデル (MTM; Mixture of Topic Models) を提案する．提案法の概要図を図 1 に示す．提案法では，各ユーザのツイート集合 $\boldsymbol{W}_u = \{\boldsymbol{W}_{us}\}_{s=1}^{D_u}$ を複数のクラスタに分割する．D_u はユーザ u のツイート数を表す．そして，同一のクラスタに割り当てられたツイート集合を擬似的に 1 文書とみなし，クラスタ毎にトピック分布を推定する．従来法 [Blei 03] では 1 文書が 1 つのトピック分布をもつが，提案法ではクラスタ毎に 1 つのトピック分布をもつため，複数のトピック分布の混合として表現できる．

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 2,

1. For each topic $k = 1, \ldots, K$
   a. draw $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta})$
2. For each user $u = 1, \ldots, U$
   a. draw $\boldsymbol{\pi}_u \sim \text{SBP}(\gamma)$
   b. For each cluster $\ell = 1, \ldots, L_u$
      i. draw $\boldsymbol{\theta}_{u\ell} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
   c. For each tweet $s = 1, \ldots, D_u$
      i. draw $y_{us} \sim \text{Categorical}(\boldsymbol{\pi}_u)$
      ii. For each word $n = 1, \ldots, N_{us}$
         A. draw $z_{usn} \sim \text{Categorical}(\boldsymbol{\theta}_{uy_{us}})$
         B. draw $w_{usn} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{usn}})$

図 2: 提案モデルの生成過程

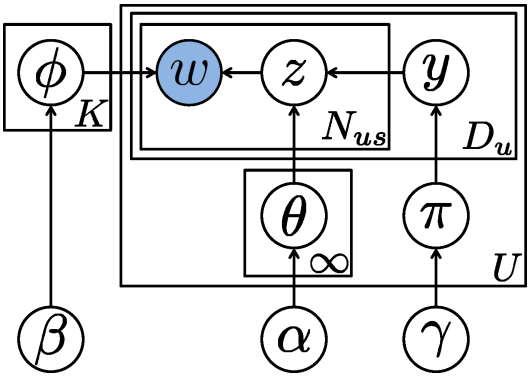

図 3: 提案モデルのグラフィカルモデル

図 3 に示す．提案モデルではユーザ毎にクラスタ分布 $\boldsymbol{\pi}_u (u = 1, \ldots, U)$, クラスタ毎にトピック分布 $\boldsymbol{\theta}_{u\ell} (\ell = 1, \ldots, L_u)$, およびトピック毎に単語分布 $\boldsymbol{\phi}_k (k = 1, \ldots, K)$ がある．ここで，U はユーザ数，K はトピック数を表す．また，L_u はユーザ u のクラスタ数を表す．クラスタ数を事前に設定することは困難であるため，ディリクレ過程 (DP; Dirichlet Process) を用いることにより，クラスタ数 L_u を推定する．はじめにユーザ u のクラスタ分布 $\boldsymbol{\pi}_u$ に従って s 番目のツイートにクラスタ $y_{us} \in \{1, \cdots, L_u\}$ を割り当てる．そして割り当てられたクラスタのトピック分布 $\boldsymbol{\theta}_{uy_{us}}$ に従ってそれぞれの単語にトピック $z_{usn} \in \{1, \cdots, K\}$ が割り当てられ，単語分布 $\boldsymbol{\phi}_{z_{usn}}$ に従って単語が生成される．ここで，トピック分布 $\boldsymbol{\theta}_{u\ell}$, および単語分布 $\boldsymbol{\phi}_k$ はカテゴリ分布のパラメータのため，その共役事前分布であるディリクレ分布から生成されると仮定し，ハイパーパラメータはそれぞれ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_V)$ である．V は語彙数を表す．また，クラスタ分布 $\boldsymbol{\pi}_u$ はディリクレ過程の構成法の 1 つである棒折り過程 (SBP; Stick-Breaking Process) から生成されると仮定し，集中パラメータは γ である．

先行研究 [McCallum 09] において，トピック分布のハイパーパラメータ $\boldsymbol{\alpha}$ は一様でなく，単語分布のハイパーパラメータ $\boldsymbol{\beta}$ は一様の場合に性能がよいことが確認されており，これ以降ではハイパーパラメータとして $\boldsymbol{\alpha}$ および β を用いる．

2.2 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を用い，クラスタ分布パラメータ $\boldsymbol{\Pi}$, トピック分布パラメータ $\boldsymbol{\Theta}$, および単語分布パラメータ $\boldsymbol{\Phi}$ を積分消去している．ツイート集合, トピック z の集合, クラスタ y の集合をそれぞれ $\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{Y}$ とすると，同時分布は以下のように導出できる．

$$p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{Y} | \boldsymbol{\alpha}, \beta, \gamma) = p(\boldsymbol{W} | \boldsymbol{Z}, \beta) \cdot p(\boldsymbol{Z} | \boldsymbol{Y}, \boldsymbol{\alpha}) \cdot p(\boldsymbol{Y} | \gamma) \quad (1)$$

(1) 式の第一項は

$$p(\boldsymbol{W} | \boldsymbol{Z}, \beta) = \prod_k \frac{\Gamma(\beta V)}{\Gamma(\beta)^V} \frac{\prod_v \Gamma(N_{kv} + \beta)}{\Gamma(N_k + \beta V)}, \quad (2)$$

第二項は

$$p(\boldsymbol{Z} | \boldsymbol{Y}, \boldsymbol{\alpha}) = \prod_u \prod_\ell \frac{\Gamma(\sum_{k'} \alpha_{k'})}{\prod_{k'} \Gamma(\alpha_{k'})} \frac{\prod_k \Gamma(N_{u\ell k} + \alpha_k)}{\Gamma(N_{u\ell} + \sum_{k'} \alpha_{k'})}, \quad (3)$$

第三項は

$$p(\boldsymbol{Y} | \gamma) = \prod_u \frac{\gamma^{L_u} \prod_\ell (D_{u\ell} - 1)!}{\gamma(\gamma + 1) \cdots (\gamma + D_u - 1)} \quad (4)$$

となる．ここで，N_{kv} は語彙 v にトピック k が割り当てられた単語数，$N_{u\ell k}$ はユーザ u のクラスタ ℓ でトピック k が割り当てられた単語数，$D_{u\ell}$ はユーザ u のツイート集合でクラスタ ℓ に割り当てられたツイート数である．また，$N_k = \sum_v N_{kv}$, $N_{u\ell} = \sum_k N_{u\ell k}$, $D_u = \sum_\ell D_{u\ell}$ である．

(3,4) 式から，ユーザ u の s 番目のツイートのクラスタ y_{us} のサンプリング確率は以下のように導出できる．

$$\begin{aligned} &p(y_{us} = \ell | \boldsymbol{Z}, \boldsymbol{Y}^{(u)}_{\setminus us}, \gamma, \boldsymbol{\alpha}) \\ &\propto p(y_{us} = \ell | \boldsymbol{Y}^{(u)}_{\setminus us}, \gamma) \cdot p(\boldsymbol{z}_{us} | \boldsymbol{Z}_{\setminus us}, y_{us} = \ell, \boldsymbol{Y}^{(u)}_{\setminus us}, \boldsymbol{\alpha}) \end{aligned} \quad (5)$$

(5) 式の第一項は

$$p(y_{us} = \ell | \boldsymbol{Y}^{(u)}_{\setminus us}, \gamma) = \begin{cases} \frac{D_{u\ell \setminus us}}{D_u - 1 + \gamma}, & \text{既存のクラスタ} \\ \frac{\gamma}{D_u - 1 + \gamma}, & \text{新規のクラスタ}, \end{cases} \quad (6)$$

第二項は

$$\begin{aligned} &p(\boldsymbol{z}_{us} | \boldsymbol{Z}_{\setminus us}, y_{us} = \ell, \boldsymbol{Y}^{(u)}_{\setminus us}, \boldsymbol{\alpha}) \\ &= \frac{\Gamma(N_{u\ell \setminus us} + \sum_{k'} \alpha_{k'})}{\Gamma(N_{u\ell \setminus us} + N_{us} + \sum_{k'} \alpha_{k'})} \prod_k \frac{\Gamma(N_{u\ell k \setminus us} + N_{usk} + \alpha_k)}{\Gamma(N_{u\ell k \setminus us} + \alpha_k)} \end{aligned} \quad (7)$$

となる．ここで，N_{us} はユーザ u の s 番目のツイートに含まれる単語数，N_{usk} はユーザ u の s 番目のツイートでトピック k が割り当てられた単語数である．また，$\boldsymbol{Y}^{(u)}$ はユーザ u のクラスタ集合，$\setminus us$ はユーザ u の s 番目のツイートを除いたときの数であることを表す．

(5) 式で $y_{us} = \ell$ となるとき，(2,3) 式から，ユーザ u の s 番目のツイートの n 番目の単語のトピック z_{usn} のサンプリング確率は以下のように導出できる．

$$\begin{aligned} &p(z_{usn} = k | \boldsymbol{W}, \boldsymbol{Z}_{\setminus usn}, y_{us} = \ell, \boldsymbol{\alpha}, \beta) \\ &\propto p(z_{usn} = k | \boldsymbol{Z}_{\setminus usn}, y_{us} = \ell, \boldsymbol{\alpha}) \\ &\quad \times p(w_{usn} | \boldsymbol{W}_{\setminus usn}, z_{usn} = k, \boldsymbol{Z}_{\setminus usn}, \beta) \end{aligned} \quad (8)$$

(8) 式の第一項は

$$p(z_{usn}=k|\boldsymbol{Z}_{\backslash usn},y_{us}=\ell,\boldsymbol{\alpha})=\frac{N_{u\ell k\backslash usn}+\alpha_k}{N_{u\ell}-1+\sum_{k'}\alpha_{k'}} \quad (9)$$

第二項は

$$p(w_{usn}|\boldsymbol{W}_{\backslash usn},z_{usn}=k,\boldsymbol{Z}_{\backslash usn},\beta)=\frac{N_{kw_{usn}\backslash usn}+\beta}{N_{k\backslash usn}+\beta V} \quad (10)$$

ここで，$\backslash usn$ はユーザ u の s 番目のツイートの n 番目の単語を除いたときの数であることを表す．

ハイパーパラメータ $\boldsymbol{\alpha},\beta$ は，不動点反復法による周辺同時尤度を最大化することにより推定する．$\boldsymbol{\alpha}$ および β の更新式は

$$\alpha_k^{\text{new}}=\alpha_k\frac{\sum_u\sum_\ell\big(\Psi(N_{u\ell k}+\alpha_k)-\Psi(\alpha_k)\big)}{\sum_u\sum_\ell\big(\Psi(N_{u\ell}+\sum_{k'}\alpha_{k'})-\Psi(\sum_{k'}\alpha_{k'})\big)} \quad (11)$$

$$\beta^{\text{new}}=\beta\frac{\sum_k\sum_v\big(\Psi(N_{kv}+\beta)-\Psi(\beta)\big)}{\sum_k\sum_v\big(\Psi(N_k+\beta V)-\Psi(\beta V)\big)} \quad (12)$$

となる．

上記の Collapsed ギブスサンプリングによるクラスタおよびトピックの推定を繰り返すことで，提案モデルの学習が行われる．MAP 推定により，クラスタ分布 $\boldsymbol{\pi}_u$，トピック分布 $\boldsymbol{\theta}_{u\ell}$ および単語分布 $\boldsymbol{\phi}_k$ を以下の式で求めることができる．

$$\pi_{u\ell}=\frac{D_{u\ell}}{D_u} \quad (13)$$

$$\theta_{u\ell k}=\frac{N_{u\ell k}+\alpha_k}{N_{u\ell}+\sum_{k'}\alpha_{k'}} \quad (14)$$

$$\phi_{kv}=\frac{N_{kv}+\beta}{N_k+\beta V} \quad (15)$$

3. 実験

提案法の有効性を評価するため，2014 年 9 月 1 日から 9 月 15 日までの間に収集した日本語ツイートデータを用いて実験を行った．各ツイートに対して，形態素解析を行った後に名詞だけを抽出し，ストップワードを取り除くなどの前処理を行い，ツイート数 229,150, ユーザ数 2,893, 語彙数 8,908 のデータを実験に用いた．モデルの学習には Collapsed ギブスサンプリングを用い，反復回数 100 回とした．また，モデルの評価尺度として以下に示すパープレキシティを用いた．

$$\text{perplexity}=\exp\left(-\frac{1}{N}\sum_u\log p(\boldsymbol{w}_u^{\text{test}}|\mathcal{M})\right) \quad (16)$$

ここで，N はテストデータ中の全単語数，$\boldsymbol{w}_u$ はユーザ u のツイート集合に含まれる全単語である．また，test はテストデータであること，$\mathcal{M}$ は確率モデルを表す．尤度は，以下の式で求めることができ，パープレキシティが低いほどテストデータを高い精度で予測できるよい確率モデルであることを示す．

$$p(\boldsymbol{w}_u|\mathcal{M})=\prod_s\prod_n\sum_\ell\pi_{u\ell}\sum_k\theta_{u\ell k}\phi_{kw_{usn}} \quad (17)$$

本実験では，提案法 MTM の比較手法として，1 ユーザの全ツイートを 1 文書とする手法 LDA, トピック情報ではなく単語情報によりクラスタを推定する手法 UM+LDA の予測性能を評価した．ここで，手法 LDA は (17) 式のクラスタ数 L_u が 1 に相当する．また，手法 UM+LDA におけるクラスタ y_{us} のサンプリング確率は以下のように導出できる．

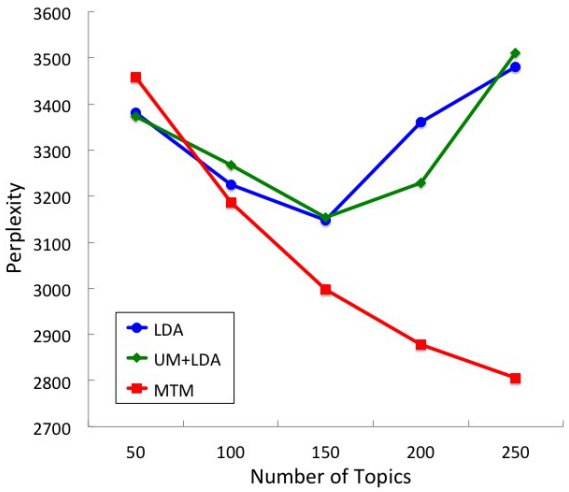

図 4: 実験結果．赤が提案法 MTM, 青が 1 ユーザの全ツイートを 1 文書とする手法 LDA, 緑が単語情報によりクラスタを推定する手法 UM+LDA を表す．また，横軸はトピック数, 縦軸はパープレキシティである．

$$\begin{aligned}&p(y_{us}=\ell|\boldsymbol{W},\boldsymbol{Y}^{(u)}_{\backslash us},\gamma,\lambda)\\&\propto p(y_{us}=\ell|\boldsymbol{Y}^{(u)}_{\backslash us},\gamma)\cdot p(\boldsymbol{w}_{us}|\boldsymbol{W}_{\backslash us},y_{us}=\ell,\boldsymbol{Y}^{(u)}_{\backslash us},\lambda)\end{aligned} \quad (18)$$

(18) 式の第一項は (6) 式と等しい．第二項は

$$\begin{aligned}&p(\boldsymbol{w}_{us}|\boldsymbol{W}_{\backslash us},y_{us}=\ell,\boldsymbol{Y}^{(u)}_{\backslash us},\lambda)\\&=\frac{\Gamma\left(N_{u\ell\backslash us}+\lambda V\right)}{\Gamma\left(N_{u\ell\backslash us}+N_{us}+\lambda V\right)}\prod_v\frac{\Gamma\left(N_{u\ell v\backslash us}+N_{usv}+\lambda\right)}{\Gamma\left(N_{u\ell v\backslash us}+\lambda\right)}\end{aligned} \quad (19)$$

となる．ここで，$N_{u\ell v}$ はユーザ u のクラスタ ℓ に含まれる語彙 v の数，N_{usv} はユーザ u の s 番目のツイートに含まれる語彙 v の数，λ はディリクレ分布のハイパーパラメータである．

ハイパーパラメータ $\boldsymbol{\alpha},\beta$ は尤度最大化により逐次推定し，集中パラメータは $\gamma=0.1$ とした．トピック数 K を 50 から 250 まで 50 ずつ変更し，それぞれツイート数 204,936 の訓練データを用いてモデルを学習し，ツイート数 24,214 のテストデータを用いてパープレキシティを求めた．結果を図 4 に示す．また，提案法 MTM の各トピック数におけるクラスタ数の平均，および手法 UM+LDA のクラスタ数の平均を表 1 にまとめる．なお，手法 UM+LDA では，クラスタ数 L_u はトピック数 K に依存しない．

図 4 より，提案法 MTM はトピック数が 100 以上のときに他の手法よりも高い精度でモデル化できていることがわかる．また，トピック数を大きくするほどモデルの精度が向上している．提案法では，単語に割り当てられたトピックの情報を用いて各ツイートにクラスタを割り当てるため，トピック数を大きくするほどクラスタの推定に用いる情報量が多くなり，より高い精度でツイート集合を分類できると考えられる．表 1 より，提案法はトピック数を大きくすることでツイート集合をより

表 1: クラスタ数の比較 (平均ツイート数は 70.8)

手法	トピック数	平均クラスタ数
MTM	50	3.69
	100	4.54
	150	4.76
	200	4.92
	250	5.06
UM+LDA	-	1.05

表 2: “bot” トピックの頻出単語上位 20 語

LDA	UM+LDA	MTM
bot	bot	ゲーム
自動	自動	bot
つぶやき	つぶやき	つぶやき
宣伝	宣伝	宣伝
設定	設定	設定
url	url	url
autotweet	autotweet	店
オートツイート	オートツイート	autotweet
活	you	オートツイート
九州	入荷	入荷
ヲタ	your	腕
マンボウ	kk	サンプル
キチガイ	km	噂
入荷	ヲタ	etc
友	マンボウ	ばか
保護	受験	ray
学芸	if	blu
東海	day	テレビ
人形	ばか	コレクション
芸人	el	奇跡

* 赤文字は各手法に共通する単語を表す.

細かいクラスタに分類していることを確認できる．また，手法 UM+LDA では，ツイート集合はほとんど同じクラスタに割り当てられている．これは，ツイートが非常に短文であるために単語の共起が起こりにくく，単語情報では適切にクラスタを推定できないためと考えられる．

次に，“bot” に関連するトピックにおける各手法の頻出単語上位 20 語を表 2 にまとめる．表 2 より，頻出上位は各手法とも「bot」や「つぶやき」,「宣伝」といった共通する単語であることがわかる．しかし，手法 LDA では,「九州」や「保護」,「人形」，手法 UM+LDA では,「you」や「受験」,「day」といった “bot” と関連なさそうな単語が含まれている．一方，提案法 MTM では,「店」や「ゲーム」,「Blu-ray」といった EC サイト関連の単語が含まれているため，運営店舗が bot により自動配信していることを推測できる．これらの結果から，提案法はよりまとまりのあるトピックを抽出できると言える．

4. おわりに

本稿では，各ユーザのツイート集合を複数のクラスタに分割し，同じクラスタに割り当てられたツイート集合を 1 文書とみなすことで，クラスタ毎に 1 つのトピック分布をもつトピックモデルを提案した．提案モデルでは，単語に割り当てられたトピックの情報を用いてクラスタを推定し，割り当てられたクラスタの情報を用いてトピックを推定している．また，クラスタ数の推定にはディリクレ過程を用い，モデルの学習には Collapsed ギブスサンプリングを用いた．日本語ツイートデータを用いた実験により，提案法が既存手法よりも高い精度でツイート集合をモデル化できることを確認した．

今後の研究では，Twitter 以外のマイクロブログのデータを用いて実験を行い，提案法の有効性を確かめる．また，提案モデルは現状，ツイートの投稿される時間的な順序を考慮していない．佐々木らは，Twitter-LDA に Twitter におけるユーザの興味と話題の時間発展を考慮したトピックモデルを提案し，提案モデルが従来モデルよりも高い精度でツイート集合をモデル化できると報告している [Sasaki 14]．そこで，提案モデルに対しても時間発展を考慮する機構を加え，提案法の有効性に対するさらなる検証，および改善を行う．

参考文献

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003)

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 50–57, New York, NY, USA (1999), ACM

[McCallum 09] McCallum, A., Mimno, D. M., and Wallach, H. M.: Rethinking LDA: Why Priors Matter, in Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A. eds., *Advances in Neural Information Processing Systems 22*, pp. 1973–1981, Curran Associates, Inc. (2009)

[Pennacchiotti 11] Pennacchiotti, M. and Popescu, A.-M.: A Machine Learning Approach to Twitter User Classification., in Adamic, L. A., Baeza-Yates, R. A., and Counts, S. eds., *ICWSM*, The AAAI Press (2011)

[Sasaki 14] Sasaki, K., Yoshikawa, T., and Furuhashi, T.: Online topic model for Twitter considering dynamics of user interests and topic trends, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1977–1985, Doha, Qatar (2014), Association for Computational Linguistics

[Weng 10] Weng, J., Lim, E.-P., Jiang, J., and He, Q.: TwitterRank: Finding Topic-sensitive Influential Twitterers, in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pp. 261–270, New York, NY, USA (2010), ACM

[Zhao 11] 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, in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pp. 338–349, Berlin, Heidelberg (2011), Springer-Verlag