Vol.2014-NL-216 No.18
Vol.2014-SLP-101 No.18
2014/5/23

ソーシャルメディア上の発信内容に基づく著者属性キーワードの推定

西山 莉紗[1,a)]　吉田 一星[1]　金山 博[1]

概要：ソーシャルメディアのユーザーの職業や趣味，家族構成といった属性を推定することは，データ収集対象とするユーザーの選択や，投稿内容の分析に有効活用できることが期待されるため，近年盛んに取り組まれている．しかし，既存手法の多くは，推定の対象とするユーザーの属性をあらかじめ数個の分類クラスとして定義する必要があるため，分析に有用なユーザー属性が前もって分からない状況では利用が難しい．このような状況を考慮し，本研究ではソーシャルメディア上の各ユーザーアカウントに提供されている著者自己紹介欄を利用して，ここに著者の属性を表す単語（著者属性キーワード）が記述される確率を推定する問題として属性推定を解くこと，そして，その推定において当該著者がソーシャルメディア上で発信している内容を利用することを提案する．筆者らは，ソーシャルメディアの著者が発信している内容からその著者属性キーワードを求めるにあたって，文書に付与されるべきタグを文書の内容に基づいて推定するタスクとの類似点に着目し，タグ推定向けのトピックモデルを応用することを試みた．本論文ではいくつかのトピックモデルを適用した結果を比較し，考察する．

Risa Nishiyama[1,a)]　Issei Yoshida[1]　Hiroshi Kanayama[1]

1. はじめに

人々の意見や行動がソーシャルメディア上で多く観測できるようになっているなか，社会現象の把握，マーケティングなどを目的とした，ソーシャルメディアのユーザーの発言の分析が盛んになっている．こういった分析は，発言を投稿したユーザーの職業や趣味，家族構成といった属性と合わせて分析することで，より価値の高い分析結果となることが期待される．例えばある政党についての評判を調査するとき，単純にこの政党がどのように好意的に捉えられているかを調べるだけでなく，どのような属性のユーザーに言及されているかを調査することで，政党の広報戦略を見直すことができる可能性がある．特に，「主婦」「社会人」といったクエリによって，それらの属性を持つユーザーを効率良く特定できれば，特定のユーザー層に関する分析を深めることができる．

ソーシャルメディア上で所望のユーザー属性情報を入手することは，一般には容易ではない．facebook [*1] などでは，勤務先，趣味，出身地などのユーザー属性が定型化されているが，そのような個別の記述形式を持たず，単一の著者自己紹介欄（author description）のみを備えているソーシャルメディアサービスも多い．ここには主に著者であるユーザーが，読者に対して自身の紹介を行う．記述の形式は一定でないものの，著者自己紹介欄の表現を利用すれば，分析に有用なユーザーの属性を特定したり，任意の属性を持つユーザーを検索できる可能性がある．

しかし，著者自己紹介欄においては，同じユーザー属性であっても異なる表現で記述されたり，開示される属性が限られていたり，注目する属性を持つユーザーを獲得することは自明ではない．例として，大学生のユーザーの投稿を分析したいとする．図 1 に示すように，さまざまな大学生のユーザーが存在するが，自己紹介欄に明示的に「大学生」と記しているのはユーザー A のみである．ユーザー B は「A 大」「ゼミ」などの大学生であることを示唆する表現を自己紹介欄に記してはいるものの，「大学生」とは記述していない．また，ユーザー C は自己紹介欄には大学生であることを示す表現を全く記しておらず，投稿された文書を参照して初めて大学生であることが推定される．

本研究の目的は，これらの課題に対し，著者自己紹介欄

[1] 日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research - Tokyo
[a)] lisa@jp.ibm.com
[*1] http://facebook.com

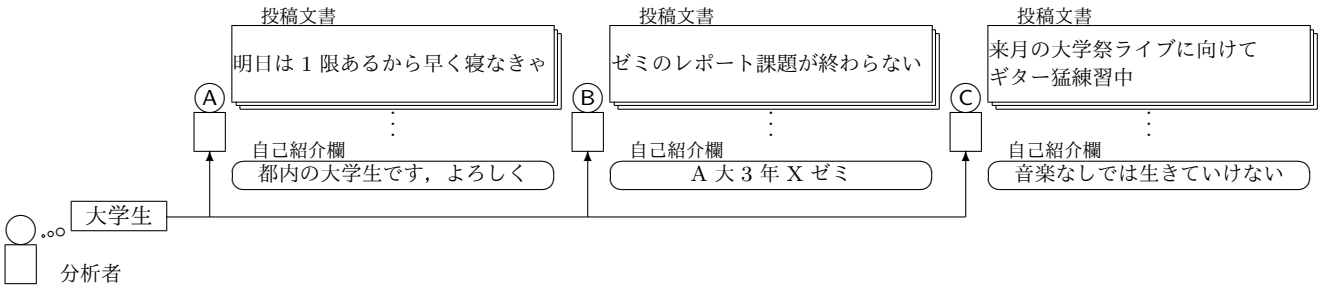


図 1 投稿文書と自己紹介欄をもとに大学生の属性を持つユーザーを取得する応用のイメージ

と投稿文書の内容を分析することにより，各ユーザーに対してその著者自己紹介欄に書かれやすいユーザー属性の表現を予測することである．解くべきタスクは，生成モデルを用いて，ユーザーに対してユーザーの属性を表す表現を推定する問題に帰着される．本稿では，使用する情報の種類や生成モデルの種類によっていくつかの異なる手法を比較検討し，それらの性質の違いについて議論する．

2. 関連研究

ソーシャルメディア上の投稿文書中の記述を利用してユーザーの属性を推定する取り組みは多数存在する．これまでの研究で推定が試みられてきた属性は主に，年代 [1]，性別 [2]，人種 [3] などである．これらの研究の多くは，SVM やパーセプトロンなどの識別モデルを利用している．例えば性別の推定であれば，あらかじめ性別が明らかになっているユーザー集合を用意しておき，各ユーザーを何らかの特徴量で表現して，分類器を学習させる．そして性別が未知のユーザーに対して分類器を適用して性別を推定する．ここで利用される特徴量として，投稿文書に記述された表現やトピックの情報が有効であることが示されている．例えば [1] はツイート中に出現する n-gram や顔文字，感情表現などを特徴量として利用しており，[3] はツイートに LDA を適用した結果を特徴量として利用している．

著者自己紹介欄にはユーザー属性に関する記述が多くみられる．そのため，上記のような分類器を利用して，例えば「ママ」という表現が著者自己紹介欄に書かれそうか否かを判別する分類器を学習させることは，理論上は可能である．しかし，著者自己紹介欄中に出現する表現は膨大にあり一貫性がないため，あらかじめ何らかの方法で取捨選択すること無く，著者自己紹介欄中に出現する全ての表現の各々に対して分類器を作成することは現実的ではない．また，1 章で述べたユーザー属性の検索のように興味の対象であるユーザー属性が動的に決まるタスクにおいては，事前にユーザー属性を取捨選択しておくこともできない．

ユーザー属性の中でも，地理上の位置情報については生成モデルを用いた推定手法が提案されている．Cheng ら [4] はユーザー u の一連のツイート $S_{tweets}(u)$ がそのユーザーのロケーション欄に書かれている都市名 i を生成する確率 $P(i|S_{tweets}(u))$ を推定する言語モデルを提案し，ロケーション欄に都市名が記述されているユーザーと，彼らが投稿したツイートを利用してこのモデルを学習させている．一方，Eisenstein ら [5] は位置と単語の関係を直接モデルするのではなく，単語を生成する隠れ変数としてトピックを仮定し，位置とトピック，トピックと単語の関係を記述したトピックモデルを提案している．しかし，いずれのモデルも位置情報の推定に特化しており，他の属性の推定にそのまま適用することは困難である．Cheng らの言語モデルでは，居住地の推定に有効なツイート中の単語（location word）をあらかじめ分類器によって特定しているが，この分類器の特徴量には，地理上の都市別に単語の投稿件数を推定するパラメータを利用しているため，位置情報の推定以外では同じ特徴量を利用することができない．また，Eisenstein らのトピックモデルにおいても，パラメータとして各トピックの地図上での平均生成位置 (base topic) と分散 (regional variance) を利用しているため，このトピックモデルを位置情報以外の属性推定にそのまま利用することができない．

3. 著者属性キーワード推定タスクと使用するデータ

本研究では，ユーザーの属性推定として次のタスクを解く．あるユーザーが過去にソーシャルメディアに投稿した一連の文書（**投稿文書**）を基にして，このユーザーがソーシャルメディア上に自身の属性を表す表現（**著者属性キーワード**，簡単のために以降は**属性キーワード**と呼ぶ）を記述する確率を推定する．

どのような表現を属性キーワードとみなすか，という点については議論があるが，本稿では著者自己紹介欄に記述される名詞を属性キーワードとみなす．なぜなら著者自己紹介欄に記述されている名詞は，そのユーザーの職業や性別，趣味，興味分野など，何らかの属性を表していることが期待されるためである．また，著者自己紹介欄には，他のユーザーから見たときに何らかの観点で有用で，検索の対象となり得る表現が記載されていることも期待される．なお，将来的にはこの属性キーワードの選択方法について更なる工夫が行われることが期待される．

また，ユーザーの投稿文書は**文書キーワード**の集合として表現される．文書キーワードとして，本研究では文書中に出現した名詞を利用する．投稿文書は，実際は複数のツイートや複数のブログ記事から構成されていたとしても，投稿ユーザーごとに全て結合して，1 件の文書として扱う．

以上のタスクを解くに当たり，本研究では文書の内容に基づいて文書のタグを推定するためのトピックモデルを応用する．次節でトピックモデルの詳細について説明する．

4. 著者属性キーワードの推定手法

本節ではまず，既存のタグ推定向けトピックモデルである Correspondence LDA[6]（Corr-LDA，**図 2** 中 (1)）について説明し，次にその拡張モデルである Noisy Annotation Topic Model[7]（NATM，図 2 中 (2)）について，Corr-LDA との差分を中心に説明する．その後，本タスクで扱う属性キーワードと投稿文書の内容の関係を考慮した新規なモデルである，Contents Description Topic Model（CDTM，図 2 中 (3)）について説明する．

本研究が扱う属性キーワードが記述される確率を推定するタスクに，タグ推定向けのトピックモデルを適用するにあたり，トピックモデル中の文書はあるユーザーの過去の投稿文書を結合したもの，タグはそのユーザーの属性キーワードとみなす．すなわち 1 ユーザーごとに 1 文書と複数の属性キーワードを持つことになる．以降は各モデルが一般に想定するデータである，文書とタグという用語を用いて説明するが，適宜本タスクの投稿文書と属性キーワードに置き換えて読まれたい．

4.1 既存のタグ推定向けトピックモデルの適用

4.1.1 Correspondence LDA (Corr-LDA)

このモデルでは，文書中の一連の単語 w と，その文書に付与された一連のタグ p にそれぞれ隠れ変数 z，a の存在を仮定する．これらの隠れ変数は一般にトピックと呼ばれる．以降のモデルの説明中に出現する各変数についての説明は**表 1** に示されている．そして，各単語とタグが以下の過程を経て生成されたものだと仮定する．なお，Dirichlet(x) は x をハイパーパラメータとするディリクレ分布，Multinomial($\boldsymbol{x}$) は確率 $\boldsymbol{x}$ に基づく多項分布を表す：

(1) 各トピック $k = 1, \ldots, K$ について以下を繰り返す：
  (a) 単語生成確率を生成する
    $\boldsymbol{\varphi}_k \sim \text{Dirichlet}(\beta)$
  (b) タグ生成確率を生成する
    $\boldsymbol{\psi}_k \sim \text{Dirichlet}(\gamma)$
(2) 各文書 $u = 1, \ldots, U$ について以下を繰り返す：
  (a) 単語トピック生成確率を生成する
    $\boldsymbol{\theta}_u \sim \text{Dirichlet}(\alpha)$
  (b) 文書 u 中の各単語 $n = 1, \ldots, N_u$ について以下を繰り返す：
    (i) 単語トピックを生成する
      $z_n \sim \text{Multinomial}(\boldsymbol{\theta}_u)$
    (ii) 単語を生成する
      $w_n \sim \text{Multinomial}(\boldsymbol{\varphi}_{z_n})$
  (c) 文書 u に付与された各タグ $m = 1, \ldots, M_u$ について以下を繰り返す：
    (i) タグトピックを生成する
      $a_m \sim \text{Multinomial}\left(\left\{\frac{N_{ku}}{N_u}\right\}_{k=1}^{K}\right)$
    (ii) タグを生成する
      $p_m \sim \text{Multinomial}(\boldsymbol{\psi}_{a_m})$

上記のように，このモデルではタグのトピックは，文書中の単語に割り当てられたトピックの割合 $\{\frac{N_{ku}}{N_u}\}_{k=1}^{K}$ に基づいて割り当てられる．このことは，全てのタグは文書中のいずれかの単語と関連を持っている，という前提に基づいている．

任意の文書 u に任意のタグ p が付与される確率 $P(p|u)$ は，$P(p|u) = \sum_{k=0}^{K} \boldsymbol{\theta}_{uk} \boldsymbol{\psi}_{kp}$ によって求められる．このモデルを適用するタスクである属性キーワードが記述される確率の推定にあたっても，この式を用いて属性キーワードの生成確率を推定する．

$P(p|u)$ の算出にあたっては，モデル中の $\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\psi}$ を学習用のデータから得る必要がある．推定に当たっては，Collapsed Gibbs Sampling が広く用いられている [8]．Collapsed Gibbs Sampling では，モデル中の隠れ変数である z と a の値を，式 (1) と式 (2) を用いて推定する．これを各単語，各タグについて順番に数百〜数千回繰り返す．

$$P(z_i = k|\boldsymbol{Z}_{\setminus i}, \boldsymbol{W}, \boldsymbol{P}, \boldsymbol{A}, \alpha, \beta, \gamma) \propto \left(N_{uk\setminus i} + \alpha\right) \frac{N_{kw_i\setminus i} + \beta}{\sum_{w=0}^{W} N_{kw\setminus i} + W\beta} \left(\frac{N_{ku\setminus i} + 1}{N_{ku\setminus i}}\right)^{M_{ku}} \tag{1}$$

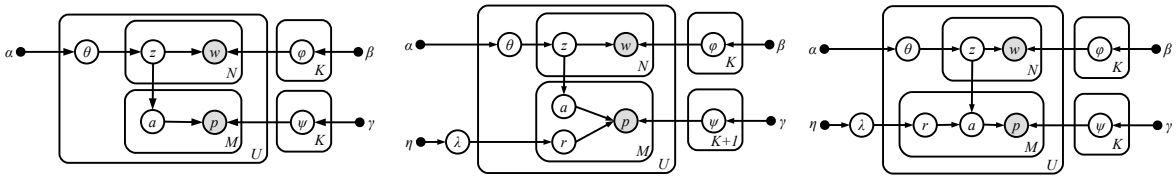
$$P(a_j = k|\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{P}, \boldsymbol{A}_{\setminus j}, \alpha, \beta, \gamma) \propto \left(M_{kp\setminus i} + \gamma\right) \frac{N_{ku}}{N_u} \tag{2}$$

Collapsed Gibbs Sampling の結果得られた z，a の値を用いて，$\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\psi}$ を以下の式で求める．

$$\boldsymbol{\theta}_u = \left(\ldots, \frac{N_{uk} + \alpha}{\sum_{k=0}^{K} N_{uk} + K\alpha}, \ldots\right) \tag{3}$$

$$\boldsymbol{\varphi}_k = \left(\ldots, \frac{N_{kw} + \beta}{\sum_{w=0}^{W} N_{kw} + W\beta}, \ldots\right) \tag{4}$$

$$\boldsymbol{\psi}_k = \left(\ldots, \frac{M_{kp} + \beta}{\sum_{p=0}^{P} M_{kp} + P\beta}, \ldots\right) \tag{5}$$



(1) Corr-LDA (2) NATM (3) CDTM

図 2 本稿で属性キーワード推定タスクに適用・比較するトピックモデルのグラフィカルモデル

表 1 記号一覧

記号	内容
U	全文書数
K	全トピック数
W	全単語異なり数
P	全タグ異なり数
N_u	文書 u に含まれる単語の数
N_{ku}	文書 u の単語のうち，トピック k を割り当てられたものの数
N_{kw}	全文書を通して，トピック k を割り当てられた単語 w の数
M_u	文書 u に付与されたタグの数
M_{ku}	文書 u のタグのうち，トピック k を割り当てられたものの数
M_{kp}	全文書を通して，トピック k を割り当てられたタグ p の数
M_0	全文書のうち，$r=0$ を割り当てられたものの数
M_1	全文書のうち，$r=1$ を割り当てられたものの数
w_n, p_m	ある文書中の n 番目の単語と，m 番目のタグ
z_n, a_m	ある文書中の n 番目の単語のトピックと，m 番目のタグのトピック
$\boldsymbol{\theta}_u$	文書 u の単語のトピック生成確率
$\boldsymbol{\varphi}_k, \boldsymbol{\psi}_k$	トピック k が生成されたときの，単語 $w=1\ldots W$，およびタグ $p=1\ldots P$ の生成確率
$\alpha, \beta, \gamma, \eta$	$\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\psi}, \lambda$ の分布を定めるハイパーパラメータ

4.1.2 Noisy Annotation Topic Model (NATM)

Corr-LDA では，全てのタグのトピックは，タグが付与された文書中の単語のいずれかに割り当てられたトピックの中から選ばれる．つまり，文書中に出現しないトピックは，その文書に付与されているタグのトピックとして割り当てられない．この前提は，文書の内容と必ず関係のあるタグが付与される状況であれば有効であるが，例えばソーシャルブックマーク上で付与される「later」（あとで読む），「reference」（参考情報）のような，内容を表さないタグが付与される状況では，文書の内容とタグの不一致を生み，推定の適合率を下げることになる．

上記に述べた，文書の内容と関連しないタグ（ノイズタグ）が付与される状況を対象として，Iwata らは Noisy Annotation Topic Model (NATM) を提案した [7]．NATM のグラフィカルモデルを図 2 中 (2) に示す．このモデルは Corr-LDA にノイズタグ (1) か否 (0) かいずれかの値を取る隠れ変数 r を付与し，さらにタグに割り当てるトピックとして，単語に付与する K 個のトピックにノイズタグ向けのトピックを加えたものである．

Corr-LDA と同様に，NATM も Collapsed Gibbs Sampling を用いて単語のトピック z，タグのトピック a，ノイズタグか否かを示すフラグ r の値を推定することができる．学習データ中の各文書に付与されている各タグについて，そのタグがノイズタグであるか否かを示す r の値を式 (6)，(7) を用いて割り当てる．

$$P(r_j=0|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{P},\boldsymbol{A},\boldsymbol{R}_{\backslash j},\alpha,\beta,\gamma,\eta) \propto \frac{M_{0\backslash j}+\eta}{M_{\backslash j}+2\eta}\frac{M_{0p_j\backslash j}+\gamma}{M_{0\backslash j}+P\gamma} \tag{6}$$

$$P(r_j=1|\boldsymbol{Z},\boldsymbol{W},\boldsymbol{P},\boldsymbol{A},\boldsymbol{R}_{\backslash j},\alpha,\beta,\gamma,\eta) \propto \frac{M_{1\backslash j}+\eta}{M_{\backslash j}+2\eta}\frac{M_{a_jp_j\backslash j}+\gamma}{M_{a_j\backslash j}+P\gamma} \tag{7}$$

割り当てた r の値を用いて，そのタグのトピックを式 (8)，(9) を用いて割り当てる．

$$P(a_j=k|r_j=0,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{P},\boldsymbol{A}_{\backslash j},\boldsymbol{R}_{\backslash j},\alpha,\beta,\gamma,\eta) \propto \frac{N_{ku}}{N_u} \tag{8}$$

$$P(a_j=k|r_j=1,\boldsymbol{Z},\boldsymbol{W},\boldsymbol{P},\boldsymbol{A}_{\backslash j},\boldsymbol{R}_{\backslash j},\alpha,\beta,\gamma,\eta) \propto \frac{M_{kp_j\backslash j}+\gamma}{M_{k\backslash j}+P\gamma}\frac{N_{ku}}{N_u} \tag{9}$$

NATM では，あらゆるトピックの文書に付与されているタグがノイズタグ ($r=0$) と判定されやすい．なぜなら式 (6)，(7) に示されるように，タグ p がノイズタグだと判断された回数 M_{0p} が，現在そのタグに割り当てられているトピック a がこれまでに割り当てられた回数 M_{ap} よりも多いと，$r=0$ が割り当てられやすくなる．あらゆるトピックの文書に付与されているタグは，M_{ap} が小さくなりやすいため，M_{0p} を大きく上回りにくくなると考えられる．従って，$r=0$ が割り当てられやすくなる．

ノイズタグと判定されたタグは，同じタグの中で一定したトピックを割り当てられにくくなる．ノイズタグと判定されたタグについては式 (8) を用いてトピック割り当てが行われるが，ノイズではないと判定されたタグに対するトピック割り当ての式 (9) と異なり，この式は他の箇所に出現した同じタグに対するトピック割り当て結果 M_{kp_j} を含んでいない．その代わりに，単語に多く付与されているトピックを優先したトピック割り当てになるため，様々なト

ピックの文書に付与されているタグについては，様々なトピックが割り当てられることになる．

NATM の上記の性質は，様々なトピックの文書に付与されるタグをノイズタグとして除外したい状況では有効に働く．しかし，本研究のように，自己紹介欄中に書かれた属性キーワードをタグとみなし，投稿文書中の単語との関連を得ようとする場合には，適切に働かない可能性がある．例えば，自己紹介欄中に属性キーワードとして「社会人」と記述しているユーザーと，そのユーザーの投稿文書中のトピックを用いて，任意のユーザーが「社会人」という属性キーワードを記述する確率を推定することを考える．このとき，投稿文書中に会社生活に関するトピックを記述しているユーザーは自己紹介欄に「社会人」と記述しやすいことが期待されるため，会社生活に関するトピックから「社会人」という属性キーワードの生成確率が高くなることが期待されるされる．しかし実際には，自己紹介欄に「社会人」と記述したユーザーが，投稿文書では趣味の話を中心に記述していることが多い．このような状況で NATM を適用した場合，「社会人」という属性キーワードがノイズタグと見なされやすくなり，一定したトピックを割り当てられにくくなることが予想される．その結果，趣味として投稿文書中に書かれていた音楽やゲームなどの，本来「社会人」とは直接関係ないはずのトピックから「社会人」が生成される確率が高くなってしまうことが予想される．

以上に述べた，属性キーワード推定タスクに NATM を適用した際の課題を解決することを目的として，筆者らは Contents Description Topic Model（CDTM）を提案する．

4.2 Contents Description Topic Model (CDTM) の提案と適用

CDTM のグラフィカルモデルを図 2 中 (3) に示す．このモデルは NATM と同様に，タグが文書の内容と関連するか否かを示す隠れ変数 r を持っている．しかし，NATM ではタグのトピック数を K から一つ増やし，r の値に応じてノイズトピックからタグを生成していたのに対し，CDTM では r はタグの生成に直接寄与せず，タグのトピック a の割り当てに寄与するモデルとなっている．

このモデルではタグのトピック a とタグは以下のように生成されると仮定する．

(1) 文書 u に付与された各タグ $m = 1, \ldots, M_u$ について以下を繰り返す：

(a) タグトピックを生成する

$$a_m \sim \begin{cases} \text{Multinomial}(\{\frac{N_{ku}}{N_u}\}_{k=1}^{K}) & (r_{um} = 1) \\ \text{Multinomial}(\frac{1}{K}) & (r_{um} = 0) \end{cases}$$

(b) タグを生成する

$$p_m \sim \text{Multinomial}(\boldsymbol{\psi}_{a_m})$$

隠れ変数 a の値を Collapsed Gibbs Sampling で推定するための条件付確率を導出すると，式 (10)，(11) のようになる．

$$P(a_j = k | r_j = 0, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{P}, \boldsymbol{A}_{\backslash j}, \boldsymbol{R}_{\backslash j}, \alpha, \beta, \gamma, \eta) \propto \frac{M_{kp_j \backslash j} + \gamma}{M_{k \backslash j} + P\gamma} \tag{10}$$

$$P(a_j = k | r_j = 1, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{P}, \boldsymbol{A}_{\backslash j}, \boldsymbol{R}_{\backslash j}, \alpha, \beta, \gamma, \eta) \propto \frac{M_{kp_j \backslash j} + \gamma}{M_{k \backslash j} + P\gamma} \frac{N_{ku}}{N_u} \tag{11}$$

式にあるように，タグのトピックの割り当てにあたっては，r の値によることなく，同じタグが他の箇所に出現した際のトピック割り当て結果 M_{kp_j} を用いる．これにより，文書と関連しないタグ ($r = 0$) であっても，文書のトピックに左右されることなく，一貫したトピックが付与されやすくなる．

各タグの r の値は，式 (12)，(13) を用いて割り当てられる．このとき，現在のタグトピック a_j が割り当てられている単語の割合 $N_{a_j u}/N_u$ が，単純に 1 をトピック数 K で割った値 $1/K$ よりも大きければ，文書と関連のあるタグ ($r = 1$) であると見なされやすくなる．

$$P(r_j = 0 | \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{P}, \boldsymbol{A}, \boldsymbol{R}_{\backslash j}, \alpha, \beta, \gamma, \eta) \propto \frac{M_{0 \backslash j} + \eta}{M_{\backslash j} + 2\eta} \frac{1}{K} \tag{12}$$

$$P(r_j = 1 | \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{P}, \boldsymbol{A}, \boldsymbol{R}_{\backslash j}, \alpha, \beta, \gamma, \eta) \propto \frac{M_{1 \backslash j} + \eta}{M_{\backslash j} + 2\eta} \frac{N_{a_j u}}{N_u} \tag{13}$$

以上に説明した 2 種類の既存のタグ推定向けトピックモデル Corr-LDA および NATM と，提案モデルである CDTM を次節の評価実験で実際のソーシャルメディアデータに適用し，性能を比較する．

5. 評価実験

5.1 実験目的

前節で説明したタグ推定向けトピックモデルを用いた著者属性キーワード推定手法は，いずれも投稿文書の内容を利用して著者属性キーワードを推定する．それに対し，ユーザーによっては初めから自己紹介欄に有効な著者属性キーワードを記載しており，著者属性キーワードを新たに推定する必要のない場合もあることが考えられる．また，所望の著者属性キーワードの記載がなくても，自己紹介欄中の他の単語を用いて著者属性キーワードを推定することが有効な場合もあり得る．例えば，「息子」「出産」などの単語を自己紹介欄に書くユーザーは，同じく自己紹介欄に「ママ」と書くことが多いため，自己紹介欄に既に記載済みの著者属性キーワードを利用して，「ママ」の生成確率を推定可能であることが考えられる．

以上を踏まえ，本実験では以下の 3 つの仮説を検証する．

仮説 1 著者自己紹介欄や投稿文書の内容を用いて著者属性キーワードを推定した方が，元々著者自己紹介欄に記載されている著者属性キーワードをそのまま著者のユーザー属性と見なした場合よりも，高い精度で著者のユーザー属性を得られる．

仮説 2 投稿文書の内容を用いて著者属性キーワードを推定した方が，著者自己紹介欄に記載された他の単語のみから著者属性キーワードを推定するよりも高い精度で著者のユーザー属性を得られる．

仮説 3 NATM や CDTM のようなタグのノイズを考慮した手法の方が，ノイズの存在を仮定しない Corr-LDA よりも高い精度で著者のユーザー属性を推定できる．

具体的な実験方法を次小節に記す．

5.2 実験方法

本実験では，ランダムに収集した 100 件の Twitter アカウントを対して人手で適切な著者属性キーワードを割り当てた正解データを作成し，仮説 (1)，(2) を検証するための 2 つのベースライン手法と，3 つのタグ推定向けトピックモデルをそれぞれ用いた提案手法の Precision(適合率)，Recall(被覆率)，F 値を比較する．

5.2.1 学習・評価用データ

自動投稿アカウント（ボットアカウント）は一貫したトピックではなく，様々なトピックのツイートを投稿するものがあることから，学習用データに含めると精度を下げる要因となる恐れがある．実際に，ツイートの内容に基づいてユーザーの GPS location を推定する既存研究 [4] では，Lee らの手法 [9] を用いて，学習用ユーザーアカウント集合から自動投稿アカウントを除外している．本研究でもこれに倣い，自動投稿アカウントを除外する．今回は簡単のために，Lee らの手法を完全に再現はせず，論文 [9] の中で有力な特徴量と報告されていた，URL を含むツイートの割合を利用した．

まず，2012 年に Twitter が提供する Streaming API を用いて収集したユーザーアカウントから，自己紹介欄の記述が存在する 18,742 アカウントを選択した．そして，その中から URL を含む過去ツイートが 3 割を上回るユーザーを自動投稿アカウントの疑いがあるとして除外した．その結果残った 17,851 ユーザーアカウントから学習用ユーザーとして 5,000 アカウント，評価用ユーザーとして学習用ユーザーとは別に，現在でもアカウントが有効な 100 ユーザーを選択した．学習用ユーザーと評価用ユーザー両方について，自己紹介欄テキストと，最大 400 件の過去ツイートを取得した．

評価用ユーザーについては，正解としてこれらのユーザーの属性を良く表していると考えられる著者属性キーワードをあらかじめ付与しておいた．具体的には，2 人の被験者が Twitter 上 [*2] で各ユーザーの自己紹介欄と最新のツイートを読み，**表 2** に示す 5 つの著者属性キーワードについて以下の 3 つの値のいずれかを付与した．

- 当てはまると言える（1）
- ほぼ確証を持って当てはまらないと言える（0）
- 当てはまるとも当てはまらないとも言えない（9）

そして，両被験者の結果を総合し，一方の被験者が 1 と判断していて，もう一方が 1 または 9 と判断したユーザー属性については当てはまる（1），それ以外のユーザー属性，すなわち両被験者とも 9 または 0 を付与したものについては，当てはまらない（0）として正解データを作成した．

表 2 正解データ作成時に付与の対象とした著者属性キーワードと，最終的に当てはまる（1）とされたユーザーの数

著者属性キーワード	正解データ中のユーザー数	属性キーワードを記述しているユーザー数
女子	52	3
社会人	43	3
大学生	20	2
アニメ	35	7
アイドル	19	1

5.2.2 ベースライン手法

推定なし (Strawman) この手法は 5.1 小節に述べた仮説 1 を検証するためのものであり，評価用ユーザーが自身の自己紹介欄に記述している著者属性キーワードをそのまま推定結果として扱う．すなわち，評価対象の著者属性キーワードがすでに自己紹介欄に記載されていれば正解，記載されていなければ不正解とする．

自己紹介欄 LDA (LDA) この手法は 5.1 小節に述べた仮説 2 を検証するためのものであり，評価用ユーザーの投稿文書は利用せず，自己紹介欄に記述している著者属性キーワードを利用して，記述されていない他の著者属性キーワードの生成確率を推定する．生成確率の推定に当たっては最も基本的な文書トピックモデルである LDA[10] を利用し，自己紹介欄テキストを文書として，文書中の単語生成確率をそのまま著者属性キーワードの生成確率とする．タグ推定向けトピックモデルと同様，LDA のパラメータ推定には Collapsed Gibbs Sampling を用いた．

5.3 実験結果と考察

まず 5.1 小節の仮説 1 を検証するにあたり，Strawman の F 値と，その他の手法の F 値を比較する．各著者属性キーワードの推定精度を**図 3** と**図 4** に示す．いずれの著者属性キーワード，いずれの推定手法であっても，Strawman の F 値を上回っており，仮説 1 が成り立っていることが分かる．

続いて仮説 2 を検証するにあたり，LDA の F 値と，Corr-LDA，NATM，そして CDTM の F 値を比較する．図 4 中

[*2] http://twitter.com

の (4) アニメおよび (5) アイドルを見ると，LDA とその他の手法の間に F 値の差がほとんど無いことがわかる．「アニメ」については，自己紹介欄に「アニメ」と合わせて具体的な作品名や声優名などの，アニメと関連する他の著者属性キーワードが書かれやすいことから，投稿文書を用いた場合と同程度の精度で「アニメ」の生成確率を推定できてしまうためだと考えられる．また，(5) アイドルについては，今回の正解データ作成時に付与の対象とした 5 つの著者属性キーワードの中で最も学習データ中の出現頻度が低いことから，LDA においては自己紹介欄中の他の著者属性キーワード，Corr-LDA，NATM，CDTM においては投稿文書中のトピックとの関係が十分に学習できなかったため，全ての手法で精度が低くなったことが考えられる．しかし，図 3 中の 3 つの著者属性キーワードについては，LDA の性能を Corr-LDA，NATM，CDTM が概ね上回っており，仮説 2 が成り立っていることが分かる．

最後に仮説 3 を検証するにあたり，Corr-LDA，NATM の F 値と，CDTM の F 値を比較した．図 3，図 4 中の 5 つの著者属性キーワードで見ると，全ての著者属性キーワードにおいて CDTM が最も良い F 値を出しているのではない．このことから，仮説 3 は必ず成り立つものではなく，推定しようとする著者属性キーワードの種類や性質によって異なると考えられる．

CDTM が最も高い F 値を出しているのは，図 3 中の (2) 社会人である．同図を見ると，この F 値の高さは高い適合率によってもたらされていることが分かる．これは，自己紹介欄に「社会人」と記載している学習用ユーザーが投稿文書に「社会人」とは直接関係がないトピックを記述している状況で，CDTM が「社会人」を文書と関連のないタグだと見なすことができたためだと期待される．CDTM が文書のトピックと関係なく「社会人」に対して一貫したトピックを割り当てようとした結果，自己紹介欄に「社会人」と記載したユーザーが偶然投稿文書に記載した関係のないトピックと「社会人」を関連付けられ，関係のないトピックから「社会人」が生成されることを避けることができた結果，適合率が向上したと考えられる．ただし，より正確には同様の性質を持つ他の著者属性キーワードを用いて検証する必要がある．

図 3 中の 3 つの著者属性キーワードで比較すると，CDTM は Corr-LDA および NATM と比較して，安定した F 値を出していることが分かる．Corr-LDA は図 3 中の (1) 女子の推定においては最も高い F 値（約 0.6）を出しているものの，(2) 社会人の推定では 10 ポイントほど下がっており（約 0.5），(3) 大学生の推定では他の 2 手法と同程度の F 値（約 0.6）となっている．また，NATM は (1) 女子の推定において，他の著者属性キーワードと比較して低い F 値（0.5 程度）を出している．これに対し，CDTM は一定して 0.6 から 0.7 の F 値を出していることが分かる．これも他の著者属性キーワードを用いてより広く検証する必要があるが，CDTM は投稿文書の内容と必ずしも関連しない著者属性キーワードを扱う本タスクにおいて安定した性能を出すことができる可能性がある．

6. おわりに

本論文ではこれまで分類器を用いて実現されてきたソーシャルメディアのユーザー属性推定を，自己紹介欄中の著者属性キーワードの生成確率を推定する問題として解くことを提案し，推定手法として，文書に付与されるべきタグを推定するための既存のトピックモデルである Corr-LDA および NATM を用いること，そして，既存のトピックモデルに本タスク向けの改良を加えた新しいモデルである CDTM を用いることを提案した．実際に Twitter ユーザーが記述した自己紹介欄と投稿文書を利用した検証実験の結果，タグ推定向けのトピックモデルを用いることで，単純に自己紹介欄の記述を利用するよりも高い精度で，ユーザーにふさわしい著者属性キーワードが推定できること，そして，特定の性質を持つ著者属性キーワードに対しては CDTM を用いることで，より高い精度で推定できる見通しがあることを示した．

今後の課題としては，より多くの著者属性キーワードを用いた精度評価実験の実施と，今回提案した推定手法による著者属性情報を用いたソーシャルメディア分析の有用性の検証が挙げられる．

参考文献

[1] Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying latent user attributes in twitter, *Proceedings of the 2nd international workshop on Search and mining user-generated contents (SMUC2010)*, New York, New York, USA, p. 37 (online), DOI: 10.1145/1871985.1871993 (2010).

[2] Burger, J. D., Henderson, J., Kim, G. and Zarrella, G.: Discriminating gender on Twitter, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2011)*, pp. 1301–1309 (online), available from ⟨http://dl.acm.org/citation.cfm?id=2145432.2145568⟩ (2011).

[3] Pennacchiotti, M. and Popescu, A.-M.: Democrats, republicans and starbucks afficionados: user classification in twitter, *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD2011)*, pp. 430–438 (online), DOI: 10.1145/2020408.2020477 (2011).

[4] Cheng, Z., Caverlee, J. and Lee, K.: You are where you tweet: A content-based approach to geo-locating twitter users, *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM2010)*, p. 759 (online), DOI: 10.1145/1871437.1871535 (2010).

[5] Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P.: A latent variable model for geographic lexical variation, *Proceedings of the 2010 Conference on*

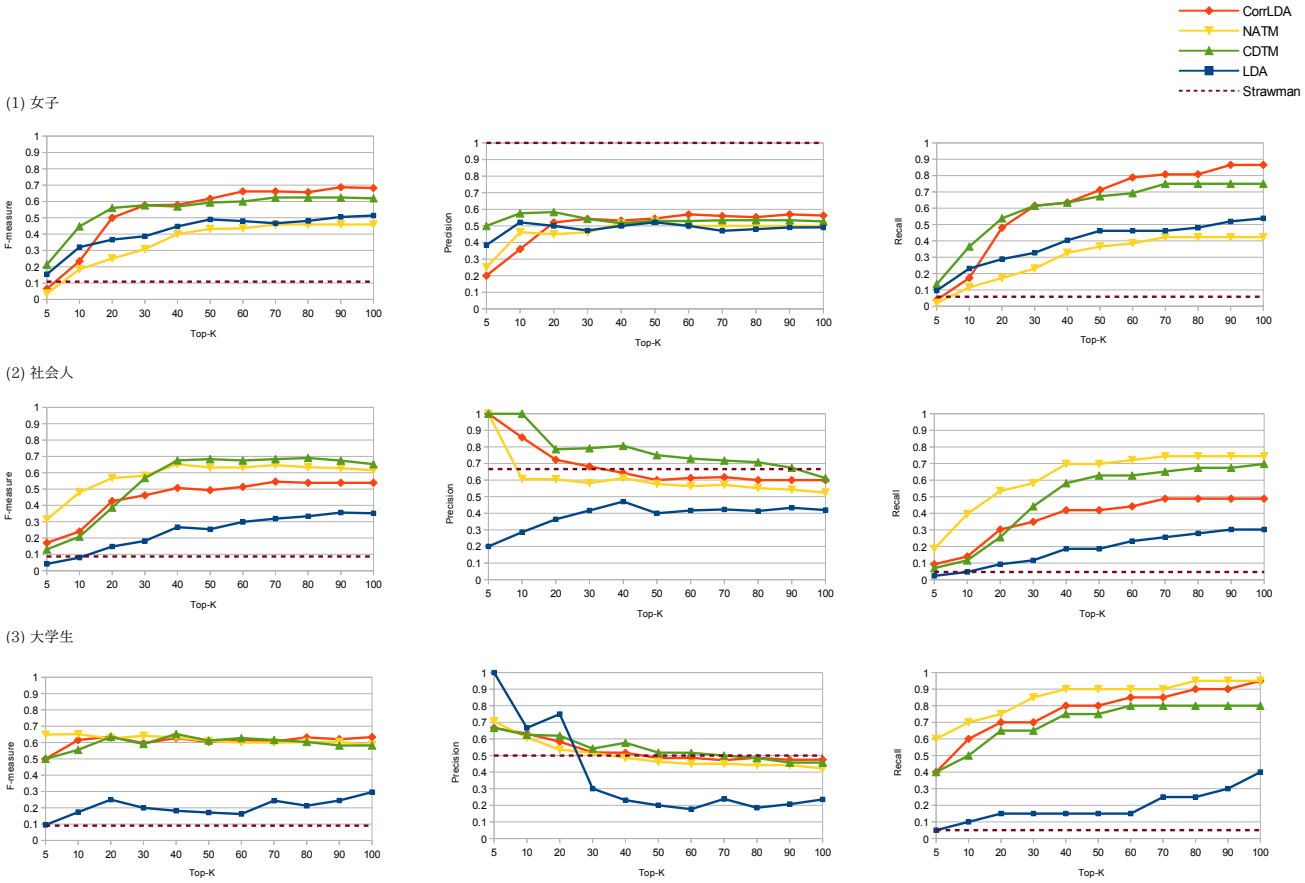


図 3 F 値 (F-measure，左列)，適合率 (Precision，中列)，再現率 (Recall，右列) 比較結果
（著者属性キーワード： 女子，社会人，大学生）

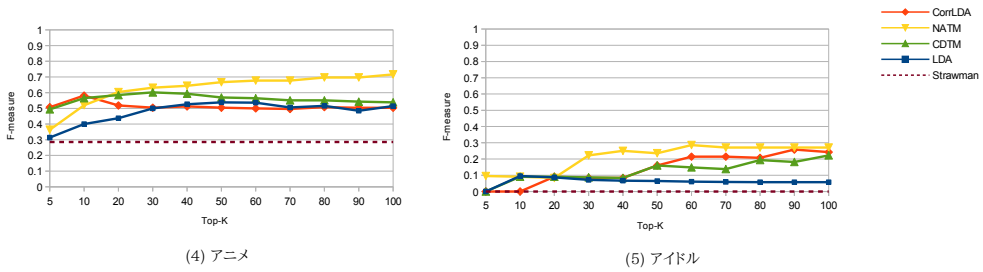


図 4 F 値比較結果 （著者属性キーワード： アニメ，アイドル）

Empirical Methods in Natural Language Processing (EMNLP2010), pp. 1277–1287 (online), available from ⟨http://dl.acm.org/citation.cfm?id=1870658.1870782⟩ (2010).

[6] Blei, D. M. and Jordan, M. I.: Modeling annotated data, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, ACM, pp. 127–134 (2003).

[7] Iwata, T., Yamada, T. and Ueda, N.: Modeling Noisy Annotated Data with Application to Social Annotation, *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 25, No. 7, pp. 1601–1613 (online), DOI: 10.1109/TKDE.2012.96 (2013).

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

[9] Lee, K., Caverlee, J. and Webb, S.: Uncovering Social Spammers: Social Honeypots + Machine Learning, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, New York, NY, USA, ACM, pp. 435–442 (online), DOI: 10.1145/1835449.1835522 (2010).

[10] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (online), available from ⟨http://dl.acm.org/citation.cfm?id=944919.944937⟩ (2003).