# トピックモデルに基づく未観測なカテゴリを含む文書集合の自動分類手法の提案

情報数理応用研究　　5215C025-9　山本祐生
指導教員　後藤正幸

# A proposal of the classification method for documents with unobserved categories based on topic model

YAMAMOTO Yusei

## 1　研究背景・目的

近年，情報技術の発達により，蓄積された膨大な文書集合を自動分類する手法の重要性が高まっている．文書分類では，予めカテゴリ情報が与えられた学習用の文書集合（以下，学習用文書）を用いて何らかの識別規則を構築し，得られた識別規則に従って分類対象であるカテゴリ情報が未知な新規文書（以下，分類対象文書）を最も所属する可能性の高いカテゴリへ分類する．通常，分類対象文書は予め学習用文書内で観測された既存のカテゴリのいずれかに分類されるため，分類対象文書に学習用文書内に存在しない未観測なカテゴリ（以下，未観測カテゴリ）に属する文書が含まれる場合，うまく分類が行えないという問題がある．そのため，新たに与えられる入力文書が学習用文書で観測された既存のカテゴリのいずれかに所属するのか，あるいは学習用文書内に存在しない未観測カテゴリに所属するのかを自動的に判別する手法が望まれる．

上記の解決策として，文書が潜在カテゴリ毎に異なる単一の確率分布によって生成されるという仮定に基づく混合 Polya 分布 [1] を用いて，未観測カテゴリを含む文書集合の生成確率をモデル化する手法が提案されている [2], [3]．この従来手法では，半教師あり学習の枠組みを援用し，学習用文書と分類対象文書の両方を用いて学習することで，未観測カテゴリに所属する文書を推定する．なかでも，従来手法 [3] はカテゴリ内に複数の潜在トピックが存在することを許容する．この手法は，カテゴリ内の潜在トピックの数だけ Polya 分布を用意するため，モデルの表現能力が増し，カテゴリ毎に 1 つの Polya 分布を仮定する従来手法 [2] と比較して優れた分類精度を示すことが確認されている．しかしこのモデルでは，各文書に対して 1 つの潜在トピックを仮定したシングルトピックモデルによって，文書集合の生成過程を表現している．

一般に新聞記事等の文書では，例えば,「東京五輪」に関する記事には「陸上」や「ボート」,「エンブレム」,「競技場」等のトピックが存在するように，1 つの文書内においても複数の潜在トピックが混在していると考えた方が自然である．しかし，従来手法 [2], [3] ではシングルトピックモデルを用いて文書をモデル化しているため，これらの複数の潜在トピックの混合を考慮できていない．すなわち，1 つの文書での複数の潜在トピックの存在を許容するマルチトピックモデルの導入により，さらなる分類精度改善の余地がある．

そこで本研究では，著者情報が付与された文書集合を対象としたマルチトピックモデルである Author Topic モデル（以下，AT モデル）を上記の問題へ援用する [4]．AT モデルは，著者情報の付与された文書集合を学習することで，各著者が嗜好する潜在トピックをモデル化する手法である．本研究では，この AT モデルを著者ではなくカテゴリに対して仮定し，カテゴリ毎に出現するトピックの特徴を学習することで，文書分類問題に適用する．さらに，新聞記事データを用いた分類実験により，従来手法と比較して提案手法の分類精度が向上することを示す．加えて，提案手法の活用法の 1 つとして，各トピックからの特徴語抽出により，カテゴリの内容理解が可能であることを示す．

## 2　準備

### 2.1　問題設定

予めカテゴリ情報が与えられた $N_L$ 件からなる学習用文書の集合を $\mathcal{D}_L = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{N_L}\}$ とし，これらの文書は $K$ 個の既存カテゴリ $c_1, c_2, \cdots, c_K$ のいずれかに所属するものとする．また，カテゴリ情報が未知な $N_T$ 件からなる分類対象文書の集合を $\mathcal{D}_T = \{\boldsymbol{x}_{N_L+1}, \boldsymbol{x}_{N_L+2}, \cdots, \boldsymbol{x}_{N_L+N_T}\}$ とする．これらの分類対象文書は既存カテゴリに加えて新たに出現する未観測カテゴリ $c_{K+1}$ を含むカテゴリ集合 $\mathcal{C} = \{c_1, c_2, \cdots, c_K, c_{K+1}\}$ のいずれかに所属するものとする．

通常の文書分類問題では，学習用文書 $\mathcal{D}_L$ を用いて識別規則の学習を行い，新たに与えられる分類対象文書 $\boldsymbol{x}_{new} \in \mathcal{D}_T$ をカテゴリへの事後確率が最大となるカテゴリ $\hat{c}$ へ分類する．

$$\begin{aligned} \hat{c} &= \arg\max_{c_k} P(c_k \mid \boldsymbol{x}_{new}) \\ &= \arg\max_{c_k} P(c_k) P(\boldsymbol{x}_{new} \mid c_k) \end{aligned} \tag{1}$$

### 2.2　記号の定義

文書の特徴量は形態素解析によって分割された各単語の出現頻度で構成されたベクトルで表される．すなわち，全文書に含まれる総異なり単語の集合を $\mathcal{W} = \{w_1, w_2, \cdots, w_V\}$ とすれば，第 $d$ 文書の特徴量は $\boldsymbol{x}_d = (x_{d1}, x_{d2}, \cdots, x_{dV})$ と表される．ただし，$w_v$ は第 $v$ 単語，$x_{dv}$ は第 $d$ 文書中に含まれる単語 $w_v$ の出現頻度とする．

### 2.3　混合 Polya 分布 [1]

文書が潜在カテゴリ毎に異なる単一の Polya 分布から生成されるという仮定に基づいた混合 Polya 分布による文書分類の手法が貞光らによって提案されている [1]．いま，混合数を $M$，混合比を $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$, $m$ 番目の Polya 分布 $P_{Polya}(\boldsymbol{x}_d; \boldsymbol{\alpha}_m)$ のパラメータを $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \cdots, \alpha_{mV})$ とすると，文書 $\boldsymbol{x}_d$ の生成確率は混合 Polya 分布によって以下のように表される．

$$\begin{aligned} P_{PM}(\boldsymbol{x}_d; \boldsymbol{\lambda}, \boldsymbol{\alpha}) &= \sum_{m=1}^{M} \lambda_m P_{Polya}(\boldsymbol{x}_d; \boldsymbol{\alpha}_m) \\ &= \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\alpha_m)}{\Gamma(\alpha_m + x_d)} \prod_{v=1}^{V} \frac{\Gamma(x_{dv} + \alpha_{mv})}{\Gamma(\alpha_{mv})} \end{aligned} \tag{2}$$

ただし，$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_M)$，$\alpha_m = \sum_{v=1}^{V} \alpha_{mv}$，$x_d = \sum_{v=1}^{V} x_{dv}$ とする．

## 3　従来研究

従来手法 [2], [3] では，半教師ありの学習の枠組みを援用し，学習用文書 $\mathcal{D}_L$ と分類対象文書 $\mathcal{D}_T$ の双方を用いて上記の混合 Polya 分布に基づき学習を行う．なかでも，従来手法 [3] では，カテゴリ「スポーツ」における

「サッカー」や「野球」といった，カテゴリ内の潜在トピックに着目し，それらに対して単一な Polya 分布を仮定する．すなわち，1 つのカテゴリを潜在トピックの数だけある複数の Polya 分布の混合によって表現する．そのもとで，各カテゴリに割り当てる混合 Polya 分布の混合数を $M(\geq K+1)$，混合比を $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_M)$ として混合したモデルにより，未観測カテゴリを含む文書集合に適用する．いま，カテゴリ $c_k$ 内に存在する $S$ 個の潜在トピック集合を $\mathcal{T}_k = \{t_{k1}, t_{k2}, \cdots, t_{kS}\}$，トピック $t_{ks}$ に仮定する Polya 分布を $P_{Polya}(\boldsymbol{x}_d; \boldsymbol{\alpha}_{ks})$，新たに導入する混合比を $\boldsymbol{\pi}_k = (\pi_{k1}, \pi_{k2}, \cdots, \pi_{kS})$ とする．このとき，$N_L$ 件からなる学習用文書 $\mathcal{D}_L$ と，$N_T$ 件からなる分類対象文書 $\mathcal{D}_T$ が独立に生起すると仮定すれば，対数尤度は以下の式 (3) によって表される．

$$\begin{aligned}
&\log L(\mathcal{D}_L, \mathcal{D}_T; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\
&= \log L(\mathcal{D}_L; \boldsymbol{\lambda}, \boldsymbol{\alpha}) + \log L(\mathcal{D}_T; \boldsymbol{\lambda}, \boldsymbol{\alpha}) \\
&= \sum_{d=1}^{N_L} \log \sum_{k=1}^{K} \delta_{dk} \lambda_k \sum_{s=1}^{S} \pi_{ks} P_{Polya}(\boldsymbol{x}_d; \boldsymbol{\alpha}_{ks}) \\
&\quad + \sum_{d=N_L+1}^{N_L+N_T} \log \sum_{m=1}^{M} \lambda_m \sum_{s=1}^{S} \pi_{ms} P_{Polya}(\boldsymbol{x}_d; \boldsymbol{\alpha}_{ms})
\end{aligned} \tag{3}$$

ただし，$\boldsymbol{\alpha}_{ms} = (\alpha_{ms1}, \alpha_{ms2}, \cdots, \alpha_{msV})$ は $t_{ks}$ における Polya 分布のパラメータ，$\delta_{dk}$ はインジケータ関数とし，文書 $\boldsymbol{x}_d$ がカテゴリ $c_k$ に所属するときに 1，そうでない場合に 0 をとるものとする．上式を極大化するパラメータ $\pi_{ms}, \lambda_m, \boldsymbol{\alpha}_{ms}$ は EM アルゴリズムによって学習する．

このモデルでは，$K \times S$ 個の Polya 分布が既存カテゴリ $c_1, c_2, \cdots, c_K$ をモデル化し，残りの $(M-K) \times S$ 個の Polya 分布が「未観測」というカテゴリ $c_{K+1}$ をモデル化する．このとき，混合 Polya 分布の総数 $M$ と，各混合 Polya 分布内の潜在トピック数 $S$ の設定によって，既存カテゴリと未観測カテゴリに割り当てられる Polya 分布の数が変動する．それに従って両カテゴリに対する分類精度が大きく左右されるため，文書集合に応じて適切な $M, S$ を調整する必要がある．ただし，従来手法 [3] は，潜在トピックを仮定しない $S = 1$ の場合，荒川らの手法 [2] と同一のモデルに帰着する．

## 4 提案手法

従来手法 [2], [3] では，1 つの文書に対して単一なトピックを仮定するシングルトピックモデルである混合 Polya 分布により文書の生成過程をモデル化している．しかしながら，新聞記事などの文書では，例えば，「東京五輪」という事柄に関する記事には「陸上」や「ボート」，「エンブレム」，「競技場」等のトピックが存在するように，1 つの文書内においても複数のトピックが混在しているものと考えた方が自然である．従来手法 [3] では，1 つのカテゴリ内に複数のトピックが含まれることを許容しているが，シングルトピックモデルにより文書の生成過程をモデル化しているため，最終的に各文書が持つトピックはカテゴリ内のいずれかの潜在トピックの 1 つという制約がある．そのため，上記の例のような，1 つの文書内に存在する複数の潜在トピックの混在をモデル化できていない．

そこで本研究では，1 つの文書を複数の潜在トピックの混合によって表現するマルチトピックモデルである AT モデルに着目する [4]．AT モデルは著者情報の付与された文書集合に適用可能な手法であり，与えられた著者情報に基づき文書集合を著者毎に分割し，それぞれに対して潜在トピックを表現することで，各著者が嗜好するトピックを分析できるモデルである．本研究では，この AT モデルをカテゴリ情報の付与された文書集合に適用し，各カテゴリに所属する文書内のトピックを学習することで，文書分類問題に適用可能とする．さらに，従来手法 [2], [3] と同様に，半教師あり学習の枠組みを用いて学習用文書と分類対象文書の両方から全文書集合に内在する潜在トピックを学習し，それにより，未観測カテゴリを含む文書集合をモデル化する手法を提案する．

### 4.1 記号の再定義

AT モデルを適用するため，文書の特徴量を以下のように再定義する．いま，第 $d$ 文書の特徴量を単語の出現頻度ではなく，$\boldsymbol{x}_d = x_{d1}, x_{d2}, \cdots, x_{di}, \cdots, x_{dn_d}$ のように，その文書に出現する単語系列として再定義する．ただし，$x_{di}$ は文書 $\boldsymbol{x}_d$ に出現する第 $i$ 番目の単語であり，$x_{di} \in \mathcal{W}$ である．また，$n_d$ は文書 $\boldsymbol{x}_d$ に含まれる総単語数とする．

### 4.2 Author Topic モデル [4]

$Q$ 人からなる著者集合を $\mathcal{A} = \{a_1, a_2, \cdots, a_q \cdots, a_Q\}$ とする．AT モデルは著者情報の付与された文書集合を学習し，著者 $a_q$ が嗜好するトピックを確率的に表現するモデルである．ただし，文書 $\boldsymbol{x}_d$ を構成するそれぞれの単語 $x_{di}$ に著者情報 $y_{di} \in \mathcal{A}$ が与えられており，文書 $\boldsymbol{x}_d$ の著者情報は系列 $\boldsymbol{y}_d = y_{d1}, y_{d2}, \cdots, y_{dn_d}$ として表現する．ここで，全文書集合に対し $J$ 個からなるトピック集合 $\mathcal{T}' = \{t'_1, t'_2, \cdots, t'_J\}$ を仮定する．マルチトピックモデルである AT モデルでは文書中の各単語に対して潜在的なトピックを仮定する．そこで，各単語のトピックを表す潜在変数を $\boldsymbol{z}_d = z_{d1}, z_{d2}, \cdots, z_{dn_d} (z_{di} \in \mathcal{T}')$ と定義する．いま，トピック $t'_j$ 内で各単語がそれぞれ確率的に出現することを表す多項分布を $\boldsymbol{\phi}_j = (\phi_{j1}, \phi_{j2}, \cdots, \phi_{jV})$，著者 $a_q$ に仮定する潜在トピックの確率分布を $\boldsymbol{\theta}_q = (\theta_{q1}, \theta_{q2}, \cdots, \theta_{qJ})$ とし，この $\boldsymbol{\theta}_q$ が著者 $a_q$ の各トピックへの嗜好度合を表す．ただし，両分布のパラメータ $\boldsymbol{\phi}_j, \boldsymbol{\theta}_q$ はそれぞれディリクレ事前分布から抽出されるものとする．AT モデルでは，与えられた著者情報 $\boldsymbol{y}_d$ に基づき文書を著者 $a_q$ 毎に 1 つに統合し，それぞれの文書に対して確率分布 $\boldsymbol{\theta}_q$ を割り当てる．それにより，著者 $a_q$ によって書かれた文書内に出現するトピックをモデル化する．AT モデルの学習では，潜在変数 $z_{di}$ を MCMC 法の 1 つである Gibbs Sampling によって繰り返し推定を行う．

### 4.3 提案モデル

本研究では，上記の AT モデルを援用し，未観測カテゴリを含む文書分類問題に適用可能なマルチトピックモデルを提案する．まず，学習用文書では予め与えられたカテゴリ情報を用いて，既存カテゴリ毎に潜在トピックを推定する．その後，学習用文書と分類対象文書の両方を用いて，文書中に含まれる各単語のカテゴリとトピックを同時に推定する．

いま，未観測カテゴリを含む $J_T$ 個のトピックからなる全てのトピック集合を $\mathcal{T}' = \{t'_1, t'_2, \cdots, t'_{J_L}, t'_{J_L+1}, \cdots, t'_{J_T}\}$ とする．ただし，$J_L$ は学習データ中に含まれる既存カテゴリのトピック数であり，$J_T \geq J_L + 1$ とする．すなわち，$t'_1, t'_2, \cdots, t'_{J_L}$ が既存カテゴリのトピックを表現し，$t'_{J_L+1}, t'_{J_L+2}, \cdots, t'_{J_T}$ が未観測カテゴリのトピックを表現する．また，文書 $\boldsymbol{x}_d$ がどのカテゴリに所属するかを示す指示変数を $l_d$，カテゴリ $c_k$ のトピック分布のパラメータを $\boldsymbol{\theta}_k = (\theta_{k1}, \theta_{k2}, \cdots, \theta_{kJ_T})$ とする．ただし，カテゴリを表すパラメータ $\boldsymbol{\theta}_k$ の数を $k = K+1$ とし，$\boldsymbol{\theta}_{K+1}$ は未観測カテゴリのトピック分布のパラメータを示すものとする．すなわち，$K$ 個のトピック分布のパラメータ $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_K$ が既存カテゴリ $c_1, c_2, \cdots, c_K$ をモデル化し，残りの $K+1$ 個目の $\boldsymbol{\theta}_{K+1}$ が未観測カテゴリ $c_{K+1}$ を表現する．また，各トピック $t'_j$ 毎の単語頻度分布のパラメータを $\boldsymbol{\phi}_j = (\phi_{j1}, \phi_{j2}, \cdots, \phi_{jV})$ とする．このとき，分類対象文書内の各単語 $x_{di}$ がどのカテゴリに割り当てられるかを表す潜在変数を $y_{di} \in \mathcal{C}$ とし，単語がどのトピックに割り当てられるかを表す潜在変数を $z_{di} \in \mathcal{T}'$ と

すると，文書 $\boldsymbol{x}_d$ の生成確率は以下の式 (4) のように表される．

$$p(\boldsymbol{x}_d \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \int\int \prod_{k=1}^{K+1} p(\boldsymbol{\theta}_k \mid \boldsymbol{\gamma}) \prod_{j=1}^{J_T} p(\boldsymbol{\phi}_j \mid \boldsymbol{\beta}) \times \prod_{i=1}^{n_d} \sum_{z_{di}} \sum_{y_{di}} p(x_{di} \mid \boldsymbol{\phi}_j) p(z_{di} \mid \boldsymbol{\theta}_k) p(y_{di} \mid l_d) d\boldsymbol{\theta}_k d\boldsymbol{\phi}_j \quad (4)$$

ただし，$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_{J_L}, \gamma_{J_L+1}, \cdots, \gamma_{J_T})$，$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_V)$ は，それぞれ $\boldsymbol{\theta}_k, \boldsymbol{\phi}_j$ に仮定されたディリクレ事前分布のハイパーパラメータである．

図 1 に，提案モデルのグラフィカルモデルを示す．

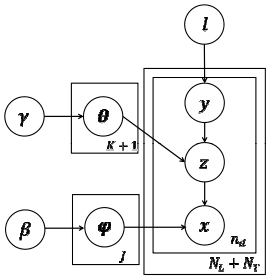


図 1. 提案モデルのグラフィカルモデル

### 4.4 モデルの学習・分類

提案手法では，従来手法 [2], [3] と同様に半教師あり学習の枠組みを援用し，学習用文書と分類対象文書の双方を用いて文書中に含まれる各単語のカテゴリとトピックを推定する．まず，学習用文書ではカテゴリ情報が与えられているため，潜在変数 $y_{di}$ については推定を行わず，学習用文書内に含まれる全単語の潜在変数 $z_{di} \in \{t'_1, t'_2, \cdots, t'_{J_L}\}$ を Gibbs Sampling によって推定する．その後，分類対象文書に関しては，潜在変数 $y_{di} \in \mathcal{C}$ と $z_{di} \in \mathcal{T}'$ を交互に繰り返し推定することで，分類対象文書内の各単語をいずれかのカテゴリおよびトピックに割り当てる

いま，全単語系列によって構成される集合 $\{\boldsymbol{x}_d\}_{d=1}^{N_L+N_T}$ から単語 $x_{di} \in \mathcal{W}$ を除いた単語集合を $\boldsymbol{x}^{-di}$，トピックを表す全潜在変数の集合 $\{z_{di}\}_{d=1,i=1}^{N_L+N_T,n_d}$ から潜在変数 $z_{di}$ を除いた集合を $\boldsymbol{z}^{-di}$，カテゴリを表す全潜在変数の集合 $\{y_{di}\}_{d=1,i=1}^{N_L+N_T,n_d}$ から潜在変数 $y_{di}$ を除いた集合を $\boldsymbol{y}^{-di}$ とすると，Gibbs Sampling によって各単語に割り当てられるカテゴリ及びトピックは以下の式 (5) を用いて推定される．

$$p(z_{di} = t'_j, y_{di} = c_k \mid x_{di} = w_v, \boldsymbol{y}^{-di}, \boldsymbol{z}^{-di}, \boldsymbol{x}^{-di}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \propto \frac{N_{kj} + \gamma}{N_{k\cdot} + \sum_{j'} \gamma_{j'}} \cdot \frac{N_{jv} + \beta_v}{N_{j\cdot} + \sum_{v'} \beta_{v'}} \quad (5)$$

ただし，$N_{kj}$ はカテゴリ $c_k$ にトピック $t'_j$ が割り当てられた回数，$N_{k\cdot}$ をカテゴリ $c_k$ に割り当てられた総トピック数，$N_{jv}$ をトピック $t'_j$ に単語 $v$ が割り当てられた回数，$N_{j\cdot}$ をトピック $t'_j$ に割り当てられた総単語数とする．式 (5) をすべての単語に対して行うことで，文書内の単語に対してカテゴリ及びトピックを割り当てる．また，ハイパーパラメータ $\gamma_j, \beta_v$ に関しては以下の式 (6), (7) を用いて推定を行う．

$$\gamma_j = \frac{\sum_{k=1}^{K+1} \Psi(N_{kj} + \bar{\gamma}_j) - (K+1)\Psi(\bar{\gamma}_j)}{\sum_{k=1}^{K+1} \Psi(N_{k\cdot} + \sum_{j'} \bar{\gamma}_{j'}) - (K+1)\Psi(\sum_{j'} \bar{\gamma}_{j'})} \quad (6)$$

$$\beta_v = \frac{\sum_{j=1}^{J} \Psi(N_{jv} + \bar{\beta}_v) - J\Psi(\bar{\beta}_v)}{\sum_{j=1}^{J} \Psi(N_{j\cdot} + \sum_{v'} \bar{\beta}_{v'}) - J\Psi(\sum_{v'} \bar{\beta}_{v'})} \quad (7)$$

ただし，$\bar{\gamma}_j$, $\bar{\beta}_v$ はそれぞれ更新前のパラメータ，$\Psi(\cdot)$ はディガンマ関数である．

これらの式 (5)–(7) をあらかじめ定めた回数まで繰り返し行い，最終的に得られた潜在変数 $y_{di}$ によって各分類対象文書内で最も多く出現するカテゴリ $\hat{c}$ へ分類する．

## 5 実験

提案手法の有効性を確認するため，実際の新聞記事データを用いた分類実験を行い，分類精度により提案手法と従来手法を比較する．また，提案手法を用いてトピック毎に特徴語を抽出した結果から，既存および未観測カテゴリの内容解析を行い，その有効性を検証する．

### 5.1 実験条件

毎日新聞 (2010 年度版) の新聞記事データ 6,000 件を用いて実験を行う．国際，政治，経済，家庭，芸能，スポーツの計 6 個のカテゴリを用い，各カテゴリ毎に 1,000 件ずつ記事をランダムに抽出した．そのうち，1 つのカテゴリを未観測カテゴリと規定したもとで残りのカテゴリを既存のカテゴリ (既存カテゴリ数 $K = 5$) とする．また，既存カテゴリからは 500 件を学習用データ，残りの 500 件をテストデータとした．一方で，未観測カテゴリにおいては抽出した 500 件をテストデータとし分類実験を行う．比較手法である混合 Polya 分布を用いた従来手法 [3] では，潜在トピック数 $S$ を $1 \sim 5$ と変動させたもとで，モデルの混合数 $M$ は $6(= K+1) \sim 15$ の範囲とし，既存カテゴリ及び未観測カテゴリの両カテゴリに対する分類精度を評価する．ただし，$S = 1$ の場合，従来手法 [3] は荒川らの手法 [2] と同一なモデルに帰着する．予備実験により，提案手法ではカテゴリ数を $6(= K+1)$ と固定したもとで，既存カテゴリに仮定するトピック数 $J_L$ を 30 と固定し，$J_T$ を $31 \sim 35$ の範囲で変化させる．また，ハイパーパラメータの初期値をそれぞれ $\boldsymbol{\gamma} = 0.01$，$\boldsymbol{\beta} = \frac{1}{V}$ と設定し，Gibbs Sampling は 2,000 回繰り返し行うものとした．既存カテゴリおよび未観測カテゴリに対する分類精度は次式によって算出する．

$$\text{分類精度} = \frac{\text{正しく分類できたテストデータ数}}{\text{テストデータ数}} \quad (8)$$

ただし，分類精度は 10 回ランダムに未観測カテゴリを抽出したもとで分類実験を行い，得られた結果に対する平均を最終的な値とする．

### 5.2 実験結果

以下，表 1, 2 に従来手法 [2], [3] による分類精度，表 3 に提案手法の分類精度を示す．従来モデル [2], [3] では，1 つの混合 Polya 分布における Polya 分布の混合数 $S$，想定する既存カテゴリと未観測カテゴリの総数 $M$ によってパフォーマンスが変化するため，これらを変えたときの分類精度を既存，未観測カテゴリに対して示している．

表 1. 従来手法 [2], [3] による既存カテゴリの分類精度

| 混合数 | $S=1$ | $S=2$ | $S=3$ | $S=4$ | $S=5$ |
|---|---|---|---|---|---|
| $M=6$ | **0.8510** | 0.8363 | 0.8241 | 0.8177 | 0.8194 |
| $M=7$ | 0.8483 | 0.8294 | 0.8104 | 0.8180 | 0.8108 |
| $M=8$ | 0.8464 | 0.8224 | 0.8165 | 0.8046 | 0.7957 |
| $M=9$ | 0.8438 | 0.8292 | 0.8027 | 0.7976 | 0.7944 |
| $M=10$ | 0.8440 | 0.8189 | 0.8048 | 0.7965 | 0.7976 |
| $M=11$ | 0.8444 | 0.8216 | 0.8024 | 0.7892 | 0.7840 |
| $M=12$ | 0.8451 | 0.8210 | 0.8012 | 0.7786 | 0.7844 |
| $M=13$ | 0.8411 | 0.8232 | 0.8017 | 0.7902 | 0.7763 |
| $M=14$ | 0.8425 | 0.8143 | 0.7983 | 0.7896 | 0.7762 |
| $M=15$ | 0.8402 | 0.8123 | 0.7966 | 0.7854 | 0.7712 |

表 1, 2 より，従来手法 [3] では潜在トピックを導入したことにより従来手法 [2] と比較して分類精度が良くなることが確認できるが，既存カテゴリと未観測カテゴリの分類精度の間にトレードオフな関係が確認できる．一方で，

表 3 より，提案手法では従来手法と比較して既存カテゴリへの分類精度を多少犠牲にするものの，全文書に仮定するトピック数 $J_T$ に関わらず，両カテゴリに対して安定して高精度な分類が行えることが確認できる．これらの結果より，従来手法 [2], [3] と比較して未観測カテゴリに対する提案手法の分類精度が向上していることが確認できる．

表 2. 従来手法 [2], [3] による未観測カテゴリの分類精度

| 混合数 | $S=1$ | $S=2$ | $S=3$ | $S=4$ | $S=5$ |
|---|---|---|---|---|---|
| $M=6$ | 0.0930 | 0.2924 | 0.3230 | 0.4454 | 0.5980 |
| $M=7$ | 0.0888 | 0.3394 | 0.4538 | 0.5354 | 0.6060 |
| $M=8$ | 0.1032 | 0.3080 | 0.4288 | 0.5810 | 0.5138 |
| $M=9$ | 0.1710 | 0.3202 | 0.4056 | 0.5024 | 0.6330 |
| $M=10$ | 0.1996 | 0.3182 | 0.5388 | 0.5170 | 0.6274 |
| $M=11$ | 0.1504 | 0.2814 | 0.4092 | 0.5816 | 0.6392 |
| $M=12$ | 0.1636 | 0.3130 | 0.3776 | 0.5278 | 0.6264 |
| $M=13$ | 0.2248 | 0.3324 | 0.4858 | 0.6054 | 0.6668 |
| $M=14$ | 0.2004 | 0.3094 | 0.4006 | 0.4232 | 0.6088 |
| $M=15$ | 0.2012 | 0.2324 | 0.4168 | 0.6278 | **0.6946** |

表 3. 提案手法による分類精度

| トピック数 | 既存カテゴリ | 未観測カテゴリ |
|---|---|---|
| $J_T$=31 | 0.7898 | 0.7411 |
| $J_T$=32 | 0.7913 | 0.7425 |
| $J_T$=33 | 0.7926 | 0.7325 |
| $J_T$=34 | 0.7899 | 0.7390 |
| $J_T$=35 | 0.7917 | 0.7346 |

### 5.3 考察

前述の通り，従来手法 [3] では未観測カテゴリに対する分類精度が向上している一方，既存カテゴリと未観測カテゴリの分類精度の間にあるトレードオフな関係を確認できる．未観測カテゴリに対しては $M$, $S$ を増やすことにより，表現能力が増したため分類精度が向上したものと考えられる．その一方で，既存カテゴリに対しては，潜在トピックを考慮しない従来手法 [2] の段階で学習用文書を用いた学習が十分に行えていたが，$S$ を増やすことでカテゴリに対して割り当てられる分布の数が過多となり，過学習が生じているため分類精度が低下したと考えられる．すなわち，従来手法 [2], [3] では，モデルの混合数 $M$ と潜在トピック数 $S$ によって既存および未観測カテゴリに対する分類精度が大きく変動するため，文書集合の性質に応じて両パラメータを調整する必要がある．それに対し，表 3 より，提案手法では文書集合全体に仮定する全トピック数 $J_T$ を変動させても両カテゴリに対して高精度な分類が行えていることが確認できる．このことから，従来手法 [3] と比較して提案手法の方が安定した分類が行えるといえる．

AT モデルに基づく提案手法では，全文書集合に含まれる単語に対してトピックを仮定し，それらを用いて各トピックの特徴を学習するため，従来手法と比較して潜在トピックの学習に十分なデータ数を確保できる．なおかつ，AT モデルでは文書中の各単語に対してカテゴリを仮定しており，1 つの文書内でカテゴリを縦断したトピックの推定が可能となるため，従来手法 [2],[3] と比較して未観測カテゴリに対する分類精度が向上したものと考えられる．

さらに，トピック毎の特徴語抽出から，本研究の提案手法がカテゴリの内容解析を可能とする手法であることを確認する．そこで，トピック数を $J_T=35$，カテゴリ「政治」を未観測カテゴリとした場合の，各カテゴリの特徴語を提案手法を用いて抽出した結果を表 4 に示す．まず，各カテゴリ $c_k$ の全トピックに対する $\theta_{kj}(j=1,2,\cdots,J_L,\cdots,J_T)$ を式 (9) により算出し，その中から上位 3 件のトピックに着目し，それらのトピックに含まれる特徴語を式 (10) で得られる $\psi_{jv}$ をソートすることによって抽出した．この結果より，提案手法では，既存カテゴリを全文書集合から推定したトピックの組み合わせで表現可能であり，さらには未観測カテゴリに対する特徴語抽出もうまく行えていることが確認できる．それにより，本研究の提案手法が実用的な手法であることが示された．

$$\theta_{kj}=\frac{N_{kj}+\gamma}{N_{k\cdot}+\sum_{j'}\gamma_{j'}} \tag{9}$$

$$\psi_{jv}=\frac{N_{jv}+\beta_v}{N_{j\cdot}+\sum_{v'}\beta_{v'}} \tag{10}$$

表 4. 提案手法による特徴語抽出

| カテゴリ | 特徴語 |
|---|---|
| 国際 | 大統領, 北朝鮮, 軍, 韓国, ロシア, イスラエル<br>欧州, ドル, 計画, 共同, 政策, アジア<br>選挙, イラン, 支持, 外交, ワシントン, 攻撃, 声明 |
| 経済 | 経営, 会長, 市場, 景気, 銀行, ユーロ, 下落<br>企業, 財政, 支援, 拡大, 方針, 危機, 市場<br>事業, 保険, 生産, 消費, 機関, 価格, ポイント |
| 家庭 | 学校, 都, 生活, 子, 心, 大学, 料理, 電話<br>女性, 会社, センター, 生活, 企業, 子供<br>毎日, 活動, 姿, 仕事, 話し, 言う, 言葉 |
| 芸能 | 作品, 教授, 現代, 文化, 人物, テーマ, テレビ<br>番組, 作家, ＮＨＫ, フジ, ＴＢＳ, 放送, ドラマ<br>本, 自分, 手, 影響, 午後, 人気, 現在 |
| スポーツ | 勝, 戦, 大会, 試合, メートル, 選手, 決勝, 優勝<br>プロ, チーム, 広島, W杯, 協会, 地区, 進出<br>五輪, 選手権, 出場, 国際, 開催, 連勝, 連覇 |
| 政治 | 党, 自民党, 菅, 鳩山, 内閣, 改革, 衆院<br>参院, 小沢, 国会, 幹事, 選挙, 記者, 衆院, 予算<br>議員, 法案, 議論, 国家, 方針, 資金, 改正, 審議 |

## 6 まとめと今後の課題

本研究では，マルチトピックモデルである AT モデルを援用して，未観測カテゴリを含む文書集合に適用した新たな自動分類手法を提案した．また，新聞記事データを用いた分類実験を行い，分類精度の観点で本研究の提案手法が従来手法と比較して優れていることを示した．さらに，トピック毎に抽出される特徴語の分析より，提案手法がカテゴリの内容解析を可能とし，実用性の観点からも優れた手法であることを示した．今後の課題として，クラスタ集合に対して事前分布を仮定したディリクレ過程 [5] を用いることにより，トピック数を文書集合に応じて自動的に決定する手法等が挙げられる．

## 参考文献


[1] 貞光九月, 三品拓也, 山本幹雄, “混合ディリクレ分布を用いたトピックに基づく言語モデル,” 電子情報通信学会論文誌, Vol. J88-D, No. 9, pp. 1771–1779, 2005.

[2] 荒川貴紀, 三川健太, 石田崇, 後藤正幸, “未観測カテゴリーを含む文書データの自動分類手法に関する研究,” 電子情報通信学会論文誌, Vol. J96-D, No. 8, pp.1956–1959, 2013.

[3] 山本祐生, 雲居玄道, 三川健太, 後藤正幸, “潜在トピックを考慮した未観測なカテゴリを含む文書集合の自動分類手法の提案, ” 日本経営工学会平成 27 年春季大会予稿集, pp.206–207, 2015

[4] M. Rosen-Zvi, T. Gliffith, M. Steyvers, and P. Smith “The author topic model for authors and documents,” *Proc. 20th Conference on Uncertainty in Artificial Intelligence*, pp.487–494, 2004.

[5] Y. W. Teh, M. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet process,” *Journal of the American Statistical Association*, Vol. 101, Issue 476, pp. 1566–1581, 2006