階層ベイズモデルによる多重トピック文書の確率的生成モデルの構築

佐藤　一誠† 　　中川　裕志††

† 東京大学大学院 情報理工学系研究科
†† 東京大学 情報基盤センター
E-mail: †sato@r.dl.itc.u-tokyo.ac.jp, ††nakagawa@dl.itc.u-tokyo.ac.jp

あらまし　近年の文書集合は，Wikipedia，Folksonomy に代表されるように多重トピックに分類される傾向にある．多重トピックによる分類とは，文書を，単一のトピックに分類するのではなく，重複を許して分類するものである．このような背景の下，多重トピックを持つ文書の確率モデルの研究が重要視されている．本研究では，多重トピック文書における確率的生成モデル:Parametric Dirichlet Mixture Model(PDMM) を提案する．PDMM は，従来の多重トピック文書の確率的生成モデル:Parametric Mixture Model(PMM) の階層ベイズ拡張と考えることができる．PMM では，各トピックにおけるモデルパラメータを等比率で混合し多重トピック文書をモデル化する．これに対し，PDMM では，各トピックにおけるモデルパラメータを，Dirichlet 分布を事前分布と仮定した混合比率で混合し，階層ベイズモデルにより多重トピック文書をモデル化する．MEDLINE コーパスを使用し，多重トピック分類タスクによって，PMM と PDMM を評価したところ，PDMM の有効性を確認した．

キーワード　多重トピック，確率モデル，階層ベイズモデル，変分ベイズ法，Dirichlet 分布，テキストマイニング

Hierarchical Bayesian Probabilistic Generative Model for Explicit Multi-Topic Document

Issei SATO† and Hiroshi NAKAGAWA††

† Graduate School of Information Sience and Technology, The University of Tokyo
†† Information Technology Center, The University of Tokyo
E-mail: †sato@r.dl.itc.u-tokyo.ac.jp, ††nakagawa@dl.itc.u-tokyo.ac.jp

Abstract　Recently, documents, such as those seen on Wikipedia and Folksonomy, have tended to be categorized into multiple topics. In this paper, we proposed a novel probabilistic generative model to deal with multiple–topic documents: Parametric Dirichlet Mixture Model(PDMM). PDMM is an expansion of an existing probabilistic generative model: Parametric Mixture Model(PMM) by hierarchical Bayes model. PMM models multiple-topic documents by mixing model parameters of each single topic with an equal mixture ratio. PDMM models multiple-topic documents by mixing model parameters of each single topic with mixture ratio following Dirichlet distribution. We evaluate PDMM and PMM by comparing F-measures using MEDLINE corpus. The evaluation showed that PDMM is more effective than PMM.

Key words　Multi-Topics, Probability Model, Hierarchical Bayes Model, Variational Bayes Method, Dirichlet distribution, Textmining

1. はじめに

近年の文書集合は，Wikipedia，Folksonomy に代表されるように多重トピックに分類される傾向にある．多重トピックによる分類とは，文書を，単一のトピックに分類するのではなく，重複を許して複数のトピックに分類するものである．このような背景の下，多重トピック文書の確率モデルの研究が重要視されている．

多重トピック文書の確率モデルとは，多重トピック文書が生成される過程を確率モデルによりモデル化したものである. 多重トピック文書の生成過程をモデル化することで、多重トピック文書の持つ固有の性質を抽出することができ，テキストマイニングへの幅広い応用が考えられる. 例えば，文書の多重トピック分類や文書中の多重トピック性を持つキーワード抽出などが

考えられる.

以下，多重トピック文書の生成過程をモデル化した確率モデルを，多重トピック文書の確率的生成モデルと呼ぶことにする.

多重トピック文書の確率的生成モデルは，主に次の 2 つに分類される. 1 つ目は，トピックとして潜在トピックを仮定するモデルである. 2 つ目は，トピックとして顕在トピックを仮定するモデルである.

1 つ目の潜在トピックを仮定する多重トピック文書モデルでは，トピックは，スポーツや経済などの明示的なトピックではなく，文書に内在するトピックを意味する. 文書を多重トピック分類する場合，潜在トピックを仮定する多重トピック文書モデルは，教師なし学習を行う. 潜在トピックを仮定する多重トピック文書モデルでは，トピック数が指定されると，各文書は，指定された数のトピックに重複を許して分類される. 作成されたトピックはその構成要素によってその特性が特徴づけられる. つまり，学習時に与えられるデータは，文書集合のみであり，文書集合をユーザの指定する数のトピックへ重複を許して分類する. 代表的なものに Latent Dirichlet Allocation(LDA) [1] [2] がある. また，潜在トピック数の自動決定を行う Hierarchical Dirichlet Process(HDP) [3] などが提案されている.

これに対し，2 つ目の顕在トピックを仮定する多重トピック文書モデルでは，トピックは，スポーツや経済などの明示的に予め与えられたトピックである. したがって，文書を多重トピック分類する場合，教師あり学習によって分類される. つまり，学習時に与えられるデータは，文書とその文書が属する多重トピックである：学習データ { 文書，< トピック 1 , トピック 2, $\cdots$ >}. このモデルでは，Parametric Mixture Model(PMM) [4] [5] が提案されている. なお，PMM が提案された当初は，PMM1,PMM2 という 2 つの PMM が提案されているが，本稿では PMM2 より有効性の高い PMM1 を PMM と呼ぶことにする.

本研究では，後者の顕在トピックを仮定するモデルを扱う. 特に，PMM のもつ問題点を改善したモデルを提案する.

以下，本稿では，第 2 節で用語説明などを行う. 第 3 節で，PMM について説明を行う. 第 4 節で，PMM の問題点を示し，提案モデルを紹介する. 第 5 節では，Medline コーパスを用いて，多重トピック分類タスクによって，提案モデルの評価を行う. さらに，考察としてトピック情報を利用した単語のランキング手法を紹介する. 第 6 節で，まとめと今後の課題を述べる.

2. 用 語 説 明

以下，本稿で使用する用語について説明する.

K を顕在トピックの総数とする. N を 1 つの文書中の単語数 (単語の重複も含む) とする. V を語彙の総数 (単語の種類数) とする. M を文書数とする. $\boldsymbol{w} = (w_1, w_2, \cdots, w_N)$ を文書ベクトルとする. 文書ベクトルは，文書中の単語を羅列した単語リストであり文書そのものを意味する. $\boldsymbol{y} = (y_1, y_2, \cdots, y_K)$ を文書 $\boldsymbol{w}$ に割り当てられたトピックベクトルとする. 多重トピッククラスとも呼ぶ. y_i は，文書 $\boldsymbol{w}$ が第 i トピックに属す (属さない) とき 1(0) の値をとる. $\boldsymbol{x} = (x_1, x_2, \cdots, x_V)$ を文書の BOW(Bag of Words) 表現とする. x_v は，文書 $\boldsymbol{w}$ 中での，単語 w_v の出現頻度を示す. w_n^v は，$w_n = (\neq)v$ のとき、1(0) の値となる変数とする. トピックの集合を $Y = \{1, 2, \cdots, K\}$ とする. トピックベクトル $\boldsymbol{y}$ において，y_i=1 である i の集合を $I_y \subset Y$ とする. I_y のすべての要素による和，積をそれぞれ $\sum_{i \in I_y}, \Pi_{i \in I_y}$ とする. $\Gamma(x)$ をガンマ関数とし，$\Psi(x)$ をプサイ (ディガンマ) 関数とする [6]. 多重トピック文書の確率的生成モデルは，多重トピッククラス $\boldsymbol{y}$ における文書 $\boldsymbol{w}$ の生成確率 ($P(\boldsymbol{w}|\boldsymbol{y})$) をモデルパラメータ Θ を用いてモデル化 ($P(\boldsymbol{w}|\boldsymbol{y}, \Theta)$) したものである. 多重トピック文書分類タスクは，トピックが既知の学習データ $D = \{(\boldsymbol{w_d}, \boldsymbol{y_d})\}_{d=1}^{M}$ をもとに，トピックが未知の文書 $\boldsymbol{w}^*$ の トピック $\boldsymbol{y}^*$ を推定する問題である.

3. Parametric Mixture Model [4] [5]

2002 年に上田，斎藤によって提案された Parametric Mixture Model(PMM) について説明する.

3.1 概　　要

PMM は，多重テキスト文書の生成過程をモデル化した確率モデルである. PMM では，各トピックにおけるモデルパラメータ (各トピックにおいて単語を生成する確率分布 (Multinomial 分布) のパラメータ) を等比率で混合し多重トピック文書における単語の生成確率をモデル化する. つまり，単一トピックにおけるモデルパラメータの重心を多重トピック文書モデルのモデルパラメータとしている. この理由は，K 個のトピックから作られるすべての多重トピックの組み合わせは $2^K - 1$ 通りあり，そのすべてに対してモデルパラメータを扱うのは現実的ではないためである.

PMM は，多重トピッククラス分類タスクにおいて NB，SVM，K-NN 法，多層ニューラルネットなどの学習器を用いた手法よりも有効な結果を出している [4] [5].

3.2 定　式　化

PMM は，文書を BOW によって表現し，以下のように定式化される.

$$P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\theta}) = \Pi_{v=1}^{V} (\varphi(w_v, \boldsymbol{y}, \boldsymbol{\theta}))^{x_v} \tag{1}$$

なお，本来，文書中の単語数 N の確率分布 $P(N)$ もモデルに組み込まれるが，分類問題には寄与しないため原論文と同様に本稿では扱わない.

$\varphi(w_v, \boldsymbol{y}, \boldsymbol{\theta})$ は，多重トピッククラス $\boldsymbol{y}$ における単語 w_v の生起確率である. トピック i に対応した混合比 $h_i(\boldsymbol{y})$ とトピック i において単語 w_v が生成する確率 θ_{iv} の線形和で表される. すなわち，以下のように定式化される.

$$\varphi(w_v, \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=0}^{K} h_i(\boldsymbol{y}) \theta_{iv} \tag{2}$$

$h_i(\boldsymbol{y})$ は以下のように定式化される.

$$h_i(\boldsymbol{y}) = 0 \quad if \quad y_i = 0 \tag{3}$$

$$h_i(\boldsymbol{y}) = \frac{y_i}{\sum_{j=1}^{K} y_j} \tag{4}$$

$$h_i(\boldsymbol{y}) \geqq 0 \tag{5}$$

$$\sum_{i=1}^{K} h_i(\boldsymbol{y}) = 1 \tag{6}$$

3.3 モデルパラメータ推定

PMM の学習アルゴリズムは，EM アルゴリズムに類似した逐次反復アルゴリズムである．このアルゴリズムを用いて，学習文書 $D = \{(\boldsymbol{w_d}, \boldsymbol{y_d})\}_{d=1}^{M}$ に対して，$\Pi_{d=1}^{M} P(\boldsymbol{w_d}|\boldsymbol{y_d}, \boldsymbol{\theta})$ を最大にするモデルパラメータ $\boldsymbol{\theta}$ を求める．本稿では，モデルパラメータ $\boldsymbol{\theta}$ の更新式のみ紹介する．各文書 $\boldsymbol{w_d}$ に対して以下の関数を導入する．

$$g_{iv}^{d}(\boldsymbol{\theta}) = \frac{h(\boldsymbol{y_d})\theta_{iv}}{\sum_{j=1}^{K} h_j(\boldsymbol{y_d})\theta_{jv}} \tag{7}$$

これを用いて，モデルパラメータ $\boldsymbol{\theta}$ の更新式は以下のようになる．

$$\theta_{iv}^{(t+1)} = \frac{1}{C}\left(\sum_{d}^{M} x_{dv} g_{iv}^{d}(\boldsymbol{\theta}^{(t)}) + \zeta - 1\right) \tag{8}$$

x_{dv} は，文書 $\boldsymbol{w_d}$ 中での，単語 w_v の出現頻度．C は，$\sum_{v=1}^{V} \theta_{iv} = 1$ となるための正規化項，ζ はスムージングパラメータで $\zeta = 2$ のとき，Laplace smoothing と呼ばれる．原論文でも $\zeta = 2$ が有効であることが示されていることから本稿でも $\zeta = 2$ を用いる．

4. 提案モデル

提案モデルについて説明する．まず，概要で PMM の問題点を述べる．次に，提案モデルの定式化行い，提案モデルによる文書の生成確率の具体的な計算手法などを説明していく．

4.1 概　　要

PMM におけるパラメータ $\boldsymbol{\theta}$(各トピックにおける各単語の生成確率) は，混合比を等比率と仮定して推定される．個々の多重トピック文書は，各トピックに対してバイアス (偏った混合比率) を持つ可能性があるが，学習は文書全体を通して行われるので，文書全体で平均するとバイアスは近似的にキャンセルされると考えられる．つまり，PMM によって学習されるモデルパラメータ $\boldsymbol{\theta}$ は文書全体におけるモデルパラメータとしては良好な推定結果となりうる．しかし，個々の文書の生成確率を計算する場合は，やはり個々の文書における各トピックに対するバイアスを考慮する必要がある．本提案モデルは，PMM におけるパラメータ $\boldsymbol{\theta}$ の混合比 $\boldsymbol{\pi}$ に対して確率分布 (Dirichlet 分布 [7]) を仮定し，PMM を階層化することで，個々の文書の各トピックに対するバイアスを考慮した生成確率を計算することができる．具体的には，文書 $\boldsymbol{w}$，トピック $\boldsymbol{y}$，モデルパラメータ $\boldsymbol{\theta}$ が与えられたもとで，$\boldsymbol{\pi}$ の確率分布 (Dirichlet 分布) の事後確率分布をベイズ推定する．つまり，$P(\boldsymbol{\pi}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\theta})$ をベイズ推定する．便宜上，提案モデルを PDMM(Parametric Dirichlet Mixture Model) と呼ぶことにする．

図 1 に，3 つのトピックが与えられた場合の各トピックのモデルパラメータの混合比 $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ を 3 次元実数空間上に表現した．混合比 $\boldsymbol{\pi}$ は，3 次元実数空間上で 2 次元シンプレックスを構成する．シンプレックス上の点が 3 つのトピックの混合比を示す．すなわち，その混合比をもつ多重トピックを示している．PMM は，混合比 $\boldsymbol{\pi}$ を等比率とみなし多重トピック文書を生成する．したがって，図 1 のシンプレックス上の重心を示す多重トピックしか生成できない．PDMM は，混合比 $\boldsymbol{\pi}$ を Dirichlet 分布に従うと仮定し多重トピック文書を生成する．したがって，図 1 のシンプレックス上のさまざまな点を示す多重トピックを生成できる．

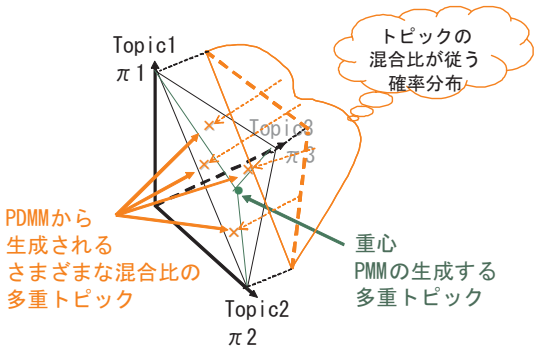

図 1　トピックシンプレックスによる多重トピック表現

4.2 定　式　化

PDMM は，PMM との対応で考えれば，以下のように定式化される．

$$P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \int P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y}) \Pi_{v=1}^{V} (\varphi(w_v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi}))^{x_v} d\boldsymbol{\pi} \tag{9}$$

$\boldsymbol{\pi}$ は $\pi_i (i \in I_y)$ を要素とするベクトルである．π_i は，トピック i のパラメータの混合比である．$\pi_i > 0, \sum_{i \in I_y} \pi_i = 1$ を満たす．π_i は，”第 i トピック度”または”第 i トピックである割合 (確率)”と考えることができる．$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})$ は，$\boldsymbol{y}$ において，$y_i = 1$ となるπ_i の事前分布である．この事前分布として Dirichlet 分布を仮定する．$\boldsymbol{\alpha}$ は，$\boldsymbol{y}$ に対応した π_i の事前分布 (ディリクレ分布) のパラメータベクトルである．すなわち，以下のように定式化される．

$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y}) = \frac{\Gamma(\sum_{i \in I_y} \alpha_i)}{\Pi_{i \in I_y} \Gamma(\alpha_i)} \Pi_{i \in I_y} \pi_i^{\alpha_i - 1} \tag{10}$$

$\varphi(w_v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi})$ は，多重トピッククラス $\boldsymbol{y}$ における単語 w_v の生起確率である．第 i トピックである割合 (確率)π_i と第 i トピックにおいて単語 w_v が生成する確率 θ_{iv} の線形和で表される．すなわち，以下のように定式化される．

$$\varphi(w_v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i \in I_y} P(y_i = 1|\boldsymbol{\pi}) P(w_v|y_i = 1, \theta) \tag{11}$$

$$= \sum_{i \in I_y} \pi_i \theta_{iv} \tag{12}$$

4.3 変分ベイズ法による個々の文書におけるトピックの混合比の事後確率分布の推定

個々の文書におけるトピックの混合比の事後確率分布 $P(\boldsymbol{\pi}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ の推定方法について説明する．事後確率分布 $P(\boldsymbol{\pi}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ は，基本的には式 (9) からベイズの定理により求めることができる．しかし，計算量が現実的ではないため変分ベイズ法 [8] [9] [10] を用いて $P(\boldsymbol{\pi}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ の近似分布を求める．以下，具体的に説明する．

式 (9)(12) より，

$$\begin{aligned} & P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) \\ = & \int P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y})\Pi_{v=1}^{V}(\sum_{i\in I_y} P(y_i=1|\boldsymbol{\pi})P(w_v|y_i=1,\theta))^{x_v}d\boldsymbol{\pi} \end{aligned} \tag{13}$$

文書表現を単語ベクトル $\boldsymbol{w}=(w_1,w_2,\cdots,w_N)$ を用いて書き換えると

$$\begin{aligned} & P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) \\ = & \int P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y})\Pi_{n=1}^{N}\sum_{i_n\in I_y} P(y_{i_n}=1|\boldsymbol{\pi})P(w_n|y_{i_n}=1,\theta)d\boldsymbol{\pi} \end{aligned} \tag{14}$$

さらに，$\sum$ と Π の順序を入れ替えると以下のように書き換えることができる．

$$\begin{aligned} & P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) \\ = & \int P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y})\sum_{i\in I_y^N}\Pi_{n=1}^{N}P(y_{i_n}=1|\boldsymbol{\pi})P(w_n|y_{i_n}=1,\theta)d\boldsymbol{\pi} \\ & (\sum_{i\in I_y^N}\equiv\sum_{i_1\in I_y}\sum_{i_2\in I_y}\cdots\sum_{i_N\in I_y}\text{とする}) \\ = & \int\sum_{i\in I_y^N} P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y})\Pi_{n=1}^{N}P(y_{i_n}=1|\boldsymbol{\pi})P(w_n|y_{i_n}=1,\theta)d\boldsymbol{\pi} \end{aligned} \tag{15}$$

$P(y_{i_n}=1|\boldsymbol{\pi})$ は，w_nにおけるトピック i の混合比率 (第 i トピックである割合 (確率)) を意味する．ここで，便宜上 $y_{i_n}=1$ を，$z_n=i$ で表現する．したがって，以下のように書き換えられる．

$$\begin{aligned} & P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) \\ = & \int\sum_{\boldsymbol{z}\in I_y^N} P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y})\Pi_{n=1}^{N}P(z_n|\boldsymbol{\pi})P(w_n|z_n,\theta)d\boldsymbol{\pi} \\ & (\sum_{\boldsymbol{z}\in I_y^N}\equiv\sum_{z_1\in I_y}\sum_{z_2\in I_y}\cdots\sum_{z_N\in I_y}\text{とする}) \end{aligned} \tag{16}$$

上記式 (16) は，新たに隠れ変数 $\boldsymbol{z}=(z_1,z_2,\cdots,z_N)$ を導入して式 (9) を表現したものと考えられる．さらに

$$P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})=\int\sum_{\boldsymbol{z}\in I_y^N}P(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})d\boldsymbol{\pi} \tag{17}$$

より式 (16) から

$$\begin{aligned} & P(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) \\ = & P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y})\Pi_{n=1}^{N}P(z_n|\boldsymbol{\pi})P(w_n|z_n,\theta) \end{aligned} \tag{18}$$

が言える．

以上の式変形により，LDA [1], [2] と同様のアプローチで変分ベイズ法を用いて $P(\boldsymbol{\pi}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ をベイズ推定することができる．

$P(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ の近似分布として $Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ を用いる．$Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ は以下のように確率変数間に因子化仮定 (近似) を仮定する．

$$Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})=Q(\boldsymbol{\pi}|\boldsymbol{\gamma})Q(\boldsymbol{z}|\boldsymbol{\phi}) \tag{19}$$

また，各々の確率分布は以下のように仮定する．

$Q(\boldsymbol{\pi}|\boldsymbol{\gamma})$ は，$\boldsymbol{\gamma}$ をパラメータとする Dirichlet 分布であり以下のように定式化される．

$$Q(\boldsymbol{\pi}|\boldsymbol{\gamma})=\frac{\Gamma(\sum_{i\in I_y}\gamma_i)}{\Pi_{i\in I_y}\Gamma(\gamma_i)}\Pi_{i\in I_y}\pi_i^{\gamma_i-1} \tag{20}$$

$Q(\boldsymbol{z}|\boldsymbol{\phi})$ は，ϕ をパラメータとする Multinomial 分布であり以下のように定式化される．

$$Q(\boldsymbol{z}|\boldsymbol{\phi})=\Pi_{n=1}^{N}Q(z_n|\boldsymbol{\phi}) \tag{21}$$

$$Q(z_n|\boldsymbol{\phi})=\Pi_{i=1}^{K}(\phi_{ni})^{z_n^i} \tag{22}$$

(z_n^iは $z_n=(\neq)i$ とき 1(0) を取る変数) (23)

$$\phi_{ni}=P(y_{i_n}=1) \tag{24}$$

以下，変分ベイズ法による近似分布 $Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$(のパラメータ $\boldsymbol{\gamma},\boldsymbol{\phi}$) の推定方法を説明する．

まず，$P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ の対数尤度を次のように式変形をする．

$$\begin{aligned} & \log P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) \\ = & \int\sum_{\boldsymbol{z}\in I_y^N}Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})d\boldsymbol{\pi}\log P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta}) && (25) \\ = & \int\sum_{\boldsymbol{z}\in I_y^N}Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})d\boldsymbol{\pi}\log\frac{P(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})}{P(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})} && (26) \\ = & \int\sum_{\boldsymbol{z}\in I_y^N}Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log\frac{P(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})}{Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})}d\boldsymbol{\pi} && (27) \\ & +\int\sum_{\boldsymbol{z}\in I_y^N}Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log\frac{Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})}{P(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})}d\boldsymbol{\pi} && (28) \end{aligned}$$

$$\mathscr{F}[Q]=\int\sum_{\boldsymbol{z}\in I_y^N}Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log\frac{P(\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})}{Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})}d\boldsymbol{\pi}$$

$$KL(Q,P)=\int\sum_{\boldsymbol{z}\in I_y^N}Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})\log\frac{Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})}{P(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})}d\boldsymbol{\pi}$$

とおくと

$$\log P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})=\mathscr{F}[Q]+KL(Q,P) \tag{29}$$

KL(Q,P) は，Kullback-Leibler Divergence と呼ばれる情報量で確率分布間の距離を測る尺度として用いられる．つまり，$Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ と $P(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ の距離を表す．したがって，式 (29) より，$\log P(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ は，$Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ に無関係な量なので，$\mathscr{F}[Q]$ を最大にする $Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ が $KL(Q,P)$ を最小にする $Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ であり，すなわち，$P(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$

の良い近似分布である．また因子化仮定(近似)により $Q(\boldsymbol{\pi}|\boldsymbol{\gamma})$ が $P(\boldsymbol{\pi}|\boldsymbol{w},\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\theta})$ の近似分布となっている．(よって，より具体的には $\mathscr{F}[Q]$ を最大にする $\boldsymbol{\gamma}$ を求めればよいことになる)

以下，$\mathscr{F}[Q]$ を最大にする $Q(\boldsymbol{\pi},\boldsymbol{z}|\boldsymbol{\gamma},\boldsymbol{\phi})$ を求める．

式 (18)(19) より $\mathscr{F}[Q]$ は以下のように書き換えられる．

$$
\begin{aligned}
&\mathscr{F}[Q] \\
&= \int Q(\boldsymbol{\pi}|\boldsymbol{\gamma}) \log P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{y}) d\boldsymbol{\theta} && (30)\\
&+ \int \sum_{\boldsymbol{z}\in I_y^N} Q(\boldsymbol{\pi}|\boldsymbol{\gamma}) Q(\boldsymbol{z}|\boldsymbol{\phi}) \log \Pi_{n=1}^N P(z_n|\boldsymbol{\pi}) d\boldsymbol{\theta} && (31)\\
&+ \sum_{\boldsymbol{z}\in I_y^N} Q(\boldsymbol{z}|\boldsymbol{\phi}) \log \Pi_{n=1}^N P(w_n|z_n,\theta) && (32)\\
&- \int Q(\boldsymbol{\pi}|\boldsymbol{\gamma}) \log Q(\boldsymbol{\pi}|\boldsymbol{\gamma}) d\boldsymbol{\theta} && (33)\\
&- \sum_{\boldsymbol{z}\in I_y^N} Q(\boldsymbol{z}|\boldsymbol{\phi}) \log Q(\boldsymbol{z}|\boldsymbol{\phi}) && (34)
\end{aligned}
$$

上記の式はそれぞれ解析的に解くことができて以下のように書き換えられる．

$$
\begin{aligned}
&\mathscr{F}[Q] \\
&= \log\Gamma(\sum_{i\in I_y}\alpha_j) - \sum_{i\in I_y}\log\Gamma(\alpha_i) \\
&\quad + \sum_{i\in I_y}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) && (35)\\
&+ \sum_{n=1}^N \sum_{i\in I_y} \phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) && (36)\\
&+ \sum_{n=1}^N \sum_{i\in I_y} \sum_{j=1}^V \phi_{ni} w_n^j \log\theta ij && (37)\\
&- \log\Gamma(\sum_{j\in I_y}\gamma_j) + \sum_{i\in I_y}\log\Gamma(\sum_{j\in I_y}\gamma_j) \\
&\quad - \sum_{i\in I_y}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) && (38)\\
&- \sum_{n=1}^N \sum_{i\in I_y} \phi_{ni}\log\phi_{ni} && (39)
\end{aligned}
$$

上記より，$\mathscr{F}[Q]$ は γ_i と ϕ_{ni} の関数となる．よって，$\mathscr{F}[Q]$ を最大とする γ_i と ϕ_{ni} を求める非線形関数の最大化問題となる．このような場合，ラグランジュ乗数法を用いることで解を求めることができる (ただし，大域的な最適解とはならない)．

以下，$\mathscr{F}[Q]$ を最大とする γ_i と ϕ_{ni} の更新式をラグランジュ乗数法により求める．まず，$\mathscr{F}[Q]$ を γ_i の関数 $\mathscr{F}[\gamma_i]$ と見なす．γ_i に関しては特に制約は無いので，$\frac{\partial\mathscr{F}[\gamma_i]}{\partial\gamma_i} = 0$ となる γ_i を求めればよい．これにより次式を得る．

$$\gamma_i = \alpha_i + \sum_{n=1}^N \phi_{ni} \qquad (i\in I_y) \tag{40}$$

次に，$\mathscr{F}[Q]$ を ϕ_{ni} の関数 $\mathscr{F}[\phi_{ni}]$ と見なす．λ をラグランジュ乗数とし $\sum_{i\in I_y}\phi_{ni} = 1$ という制約を考慮すれば，ラグランジュ関数 $L[\phi_{ni}]$ は次式となる．

$$L[\phi_{ni}] = \mathscr{F}[\phi_{ni}] + \lambda(\sum_{i\in I_y}\phi_{ni} - 1) \tag{41}$$

$\frac{\partial L[\phi_{ni}]}{\partial\phi ni} = 0$ となる ϕ_{ni} を求めると，次式を得る．

$$\phi_{ni} = \frac{\theta_{iw_n}}{C}\exp(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) \tag{42}$$

$(i\in I_y, C$ は正規化定数$)$

以上により以下の更新式を得る．

$$\gamma_i^{(t+1)} = \alpha_i + \sum_{n=1}^N \phi_{ni}^{(t)} \qquad (i\in I_y) \tag{43}$$

$$\phi_{ni}^{(t+1)} = \frac{\theta_{iw_n}}{C}\exp(\Psi(\gamma_i^{(t+1)}) - \Psi(\sum_{j\in I_y}\gamma_j^{(t+1)})) \tag{44}$$

$(i\in I_y, C$ は正規化定数$)$

上記を用いて個々の文書に固有の $\boldsymbol{\gamma},\boldsymbol{\phi}$ を推定することができる．つまり，文書 $\boldsymbol{w}$，多重トピック $\boldsymbol{y}$ が与えられたときの混合比 $\boldsymbol{\pi}$ の事後確率分布 (の近似分布) を求めることができる．なお，PDMM では，$\boldsymbol{\theta}$ は PMM で学習したものを使用する．この理由は，節 4.1 で述べたとおり文書全体としては良い推定になっていると考えたからである．また，π の事前分布 (Dirichlet 分布) のパラメータである $\boldsymbol{\alpha}$ は，すべて 1 とした．この理由は，Dirichlet 分布は，パラメータが 1 ときは一様分布になるためである．もし，π に関して事前に知識がありそれを表現したい場合は，α を調整すればよい．以下にまとめとして擬似コードを示す．

• Variational Bayes Method for PDMM————————————

function vb($\boldsymbol{w},\boldsymbol{y}$):

（1） Compute $\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\phi}^{(t+1)}$ using Eq.(43)(44)

（2） if $\| \boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^{(t)} \| < \epsilon$, $\| \boldsymbol{\phi}^{(t+1)} - \boldsymbol{\phi}^{(t)} \| < \epsilon$

（3） then return $(\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\phi}^{(t+1)})$ and halt

（4） else $t \leftarrow t+1$ and goto step (1)

————————————————————————————————

4.4 文書の生成確率の計算

PMM では，多重トピック $\boldsymbol{y}$ における文書 $\boldsymbol{w}$ の生成確率 $P(\boldsymbol{w}|\boldsymbol{y})$ を以下のように算出する．

$$P(\boldsymbol{w}|\boldsymbol{y}) = \Pi_{v=1}^V (P(w_v|\boldsymbol{y},\boldsymbol{\theta}))^{x_v} \tag{45}$$

$P(w_v|\boldsymbol{y},\boldsymbol{\theta}) = \varphi(w_v,\boldsymbol{y},\boldsymbol{\theta})$ は多重トピック $\boldsymbol{y}$ における単語 w_v の生成確率である．$\varphi(w_v,y,\theta)$ は，各トピックにおけるパラメータ $\theta_i v (i\in I_y)$ を等確率に混合したものである．

これに対し，PDMM では，各トピックにおけるパラメータ $\theta_i v (i\in I_y)$ とその混合比の事後確率分布の近似分布 $Q(\pi|\boldsymbol{\gamma})$ を用いて以下のように求める．

$$P(w_v|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\gamma}) = \int (\sum_{i\in I_y}\pi_i\theta_{iv}) Q(\boldsymbol{\pi}|\boldsymbol{\gamma}) d\boldsymbol{\pi} \tag{46}$$

上式は

$$
\begin{aligned}
P(w_v|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\gamma}) &= \int \sum_{i\in I_y}\pi_i\theta_{iv} Q(\boldsymbol{\pi},\boldsymbol{\gamma}) d\boldsymbol{\pi} && (47)\\
&= \sum_{i\in I_y}\tilde{\pi}_i\theta_{iv} && (48)
\end{aligned}
$$

$$\tilde{\pi}_i = \int \pi_i Q(\pi|\gamma) d\pi = \frac{\gamma_i}{\sum_{j \in I_y} \gamma_j} \tag{49}$$

(式 (49) は, [7] を参照のこと)

より，$Q(\pi|\gamma)$ の期待値 $\tilde{\pi}_i (i \in I_y)$ を混合比と見なしていることに相当する．つまり，多重トピック $\boldsymbol{y}$ と文書 $\boldsymbol{w}$ が与えられたとき，混合比は，変分ベイズ法により求めた事後確率分布 $Q(\boldsymbol{\pi}|\boldsymbol{\gamma})$ の平均となる．したがって，多重トピック $\boldsymbol{y}$ における文書 $\boldsymbol{w}$ の生成確率 $P(\boldsymbol{w}|\boldsymbol{y})$ は以下のようにして求める．

$$P(\boldsymbol{w}|\boldsymbol{y}) = \Pi_{v=1}^{V} (P(w_v|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}))^{x_v} \tag{50}$$

4.5 多重トピッククラス分類の予測アルゴリズム

PDMM を用いた多重トピッククラス分類の予測は，PMM と同様に，文書 $\boldsymbol{w}$ の生成確率 $P(\boldsymbol{w}|\boldsymbol{y})$ が最も高くなる多重トピッククラス $\boldsymbol{y}$ を予測することである．これは 0-1 整数最適化問題であるため PMM と同様の近似アルゴリズムを用いる．ただし，$\boldsymbol{y}$ を予測する際に混合比を変分ベイズ法で推定する点 (以下の予測アルゴリズムの擬似コードの (6) の vb($\boldsymbol{w}, \boldsymbol{y}$)) が PMM と異なる．以下に，多重トピッククラス分類の予測アルゴリズムの擬似コードを示す．

• Topics Prediction Algorithm————————————

function prediction($\boldsymbol{w}$):

（1） Initialize $S \leftarrow \{1, 2, \cdots\}, y_i \leftarrow 0$ for $i(1, 2 \cdots, K)$
（2） $v_{max} \leftarrow -\infty$
（3） while S is not empty do
（4） foreach $i \in S$ do
（5） $y_i \leftarrow 1, y_{j \in S \backslash i} \leftarrow 0$
（6） Compute $\boldsymbol{\gamma}$ by vb($\boldsymbol{w}, \boldsymbol{y}$)
（7） $v(i) \leftarrow P(\boldsymbol{w}|\boldsymbol{y})$
（8） end foreach
（9） $i^* \leftarrow$ argmax $v(i)$
（10） if $v(i^*) > v_{max}$
（11） $y_{i*} \leftarrow 1, S \leftarrow S \backslash i^*, v_{max} \leftarrow v(i^*)$
（12） else
（13） return $\boldsymbol{y}$ and halt
（14） end if
（15） end while

————————————————————————————

上記の予測アルゴリズムでは，ステップ (6) で変分ベイズ法を行うため PMM と比べて処理時間が大幅に増えてしまう．したがって，実際には，ステップ (1) において，S に含まれるトピックを以下のように予め絞り込むことで高速化を行う．まず，文書に対し、すべてのトピックを付加することで，すべてのトピックの付加を仮定した $\boldsymbol{\gamma}_{all}$ を推定する．次に，この $\boldsymbol{\gamma}_{all}$ をもとに各トピックの混合比率の大小を近似的に見積もる．$\boldsymbol{\gamma}_{all}$ の最大値 γ_{max} との比率を考えて，閾値 δ 以上であれば，付加されるトピックの候補として S に含める．S を初期化する擬似コードを以下示す．

•Initialize S Algorithm————————————————

Initialize S :

（1） $y_i \leftarrow 1$ for $i(1, 2, \cdots, K)$
（2） Compute $\boldsymbol{\gamma}_{all}$ by vb($\boldsymbol{w}, \boldsymbol{y}$)
（3） foreach $i \in \{1, 2, \cdots, K\}$
（4） if $\gamma_i / \gamma_{max} > \delta$
（5） S.push(i)
（6） end if
（7） end foreach

————————————————————————————

5. 評価実験

本提案モデルの評価実験として，多重トピック分類タスクにおける F-measure の比較を行う．

5.1 データセット

データセットとして MEDLINE コーパス(注1)を使用した．MEDLINE は生物医学分野の文献データベースである．本実験では，5000 本の英語のアブストラクトを用いた．MEDLINE には，MeSHTerm というメタデータが多重に付加されている．例えば，1 つのアブストラクトに対し Algorithms，RNA Messenger，DNA-Binding Proteins などの MeSHTerm が付加されている．これをアブストラクトの多重トピックとみなし実験を行った．ただし，トピック (MeSHTerm) は出現頻度が中頻度 (100-999) のものを用いた．この理由は，ほとんどのアブストラクトに表れる高頻度のトピックや 1 回しか表れない低頻度のトピックがあると実験結果がそのトピックに引きずられる可能性があるからである．その結果，トピック数は 88 トピックとなった．語彙数は，46,075 語．全文書に対する多重トピック文書の割合は 69.8%であった．つまり，データセット中の約 7 割の文書が多重トピックを持っていて，約 3 割の文書がシングルトピックである．多重トピック文書の 1 文書あたりの平均トピック数は 3.4 トピックであった．TreeTagger(注2)を用いて単語を base form に変換し，冠詞や be 動詞などの stop words は除去したが TF-IDF などによるフィルタリグはかけていない．

5.2 実　　験

本提案モデルの評価実験として，多重トピック分類タスクにおける F-measure の比較を行った．F-measure(F) は，適合率 (Precision:P) と再現率 (Recall:R) を用いて以下のようにして求める．

$$F = \frac{2PR}{P+R} \tag{51}$$

$$P = \frac{\text{推定トピックの中で正解した数}}{\text{推定トピックの数}} \tag{52}$$

$$R = \frac{\text{推定トピックの中で正解した数}}{\text{実際の正解トピックの数}} \tag{53}$$

F-measure が高いほど分類能力が高いことが示される．表 1 は，学習データとして 4,900 文書，テストデータとして 100 文書を用いた場合の PMM と PDMM の F-measure と 1 文書あたりのトピック予測時間 [ms] を示している．PDMM では，変分ベイズ法による予測を行うため，4.5 節の閾値 δ による候補の絞込みを行わない場合，PMM に比べて 100 倍近く速度が遅く

(注1)：http://www.nlm.nih.gov/pubs/factsheets/medline.html
(注2)：http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

表 1　トピック予測速度の比較

	PMM	PDMM (δ=0.01)	PDMM (δ=0.05)	PDMM (δ=0.1)
Time[ms]	18.1	1843	608.4	276.2
F-measure	0.320886	0.361077	0.3611698	0.3521806

なってしまう．表 1 をもとに，比較的速度が早く F-measure の高い δ=0.05 を本実験では用いる．

図 2 に，F-measure における PDMM，PMM の評価結果を示す．なお，PDMM は，モデルパラメータ $\boldsymbol{\theta}$(各トピックにおける単語の出現 (生成) 確率) の学習に PMM と同様のアルゴリズム (式 (8)) を用いるが，この $\boldsymbol{\theta}$ の学習方法の違いにより，次の 2 つのモデルも実験に加えた．$\boldsymbol{\theta}$ を Naive Bayes 法により学習するモデル **NBM** と Emprical Bayes 法により学習するモデル **EBM** である．

NBM での $\boldsymbol{\theta}$ の更新式は，$\theta_{iv} = \frac{M_{iv}+1}{C}$．$M_{iv}$ は，トピック i において単語 w_v が出現した文書数である．C は正規化項である．

EBM は，LDA [1], [2] と同様の学習アルゴリズムを用いる．**EBM** での $\boldsymbol{\theta}$ の更新式は，$\theta_{iv} = \frac{\sum_{d=1}^{M_i}\sum_{n=1}^{N_d}\phi^*_{dni}w^v_{dn}}{C}$．$M_i$ は，トピック i に分類されている文書数である．N_d は，文書 d 中の単語数である．ϕ^*_{dni} は，文書 d において変分ベイズ法により推定されたパラメータ ϕ_{ni} である．w^v_{dn} は，文書 d において，n 番目の単語が w_v だったときに 1，それ以外のとき 0 をとる変数である．LDA と同様に,θ の更新と γ，ϕ の更新を交互に行う．

図 2 の横軸は，データセット (5000 文書) に対するテストデータの割合を示す．例えば，2%の場合，4900 文書で学習し 100 文書で予測を行ったことを意味する．実験では，各テストデータの割合毎に，ランダムに 5 つのデータセットを作成し，その平均値を用いた．図 2 は，どのテストデータの割合においても，PDMM は他のモデルと比べて，F-measure が高いことを示している．よって多重トピック分類問題に対しても PDMM が有効であることがわかる．

図 3 は，データセットに対する多重トピックの文書数の割合の変化と F-measure の比較を行ったものである．図 3 の横軸は，データセットに対する多重トピック文書の文書数の割合を示す．多重トピック文書の割合が 30%の場合では，各モデル間の F-measure の違いは少ない．しかし，多重トピック文書の割合が 90%に近づくにつれ，つまり多重トピック文書が多くなるにつれ，各モデル間に F-measure の違いが出てきている．特に，PDMM は，他のモデルとの F-measure の開きが大きくなっていることから，PDMM が多重トピック文書を効果的にモデル化できていると考えられる．

5.3　考　　察

前節の実験結果から，本提案モデルは，F-measure において PMM より良好な結果を出している．PMM では，文書の生成確率を計算する際に，トピックの混合比を等比率とみなす．一方，PDMM では，文書の生成確率を計算する際に，トピックの混合比の確率分布を推定し，そこから得られる混合比を用

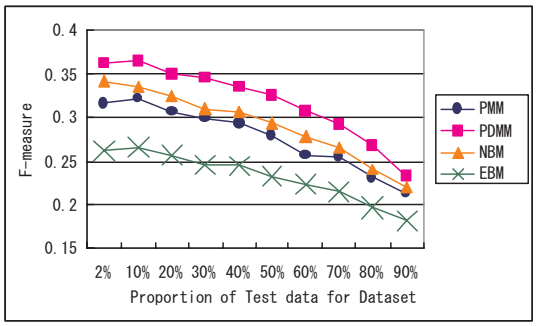

図 2　Comparison of Models with respect to F-measure

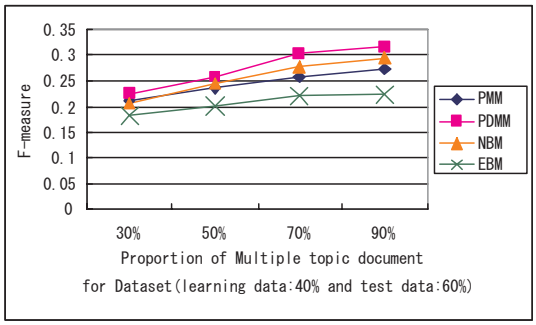

図 3　Comparison of Models with respect to F-measure changing Proportion of Multiple Topic Document for Dataset(Lerning data: 40% and Test data 60%)

いる．この混合比の推定が有意であるため上記の実験に対し良好な結果を生んでいると考えることができる．さらに，トピックの混合比の推定は，文書の特徴量の抽出の観点からも有用である可能性がある．トピックの混合比は文書に対して固有である．よって，混合比の推定は，文書の単語の出現頻度の空間 Q^V(Q : 整数, V : 語彙数) から，文書におけるトピックの割合の空間 $[0,1]^K$(K : 全トピック数) への写像と見なすことができる．一般的に語彙数はトピック数より非常に多いことから，これは次元圧縮であり多重トピック文書におけるひとつの特徴量の抽出と考えられる．これにより文書間類似度計算などへの応用が考えられる．例えば，ある文書 (アブストラクト) に対し多重トピックが [Comparative Study], [Apoptosis], [Models,Biological] であったとき，各々の推定混合比は (0.656, 0.176, 0.168) であった．この混合比がこの文書の次元圧縮された特徴量であると考えられる．

トピックの混合比の推定は第 4.3 節における変分パラメータ $\boldsymbol{\gamma}$ を用いて行われるが，この $\boldsymbol{\gamma}$ の算出に使われるもう 1 つの変分パラメータ $\boldsymbol{\phi}$ を用いることで興味深い実験結果を得ることができた．以下，具体的に説明する．

ϕ_{ni} は，単語 w_n の第 i トピックへのバイアスを意味する．したがって，以下の式により単語 w_n のトピックにおけるエントロピーを計算することができる．

$$entropy(w_n) = \sum_{i=1}^{K} \phi_{ni} \log(\phi_{ni}) \qquad (54)$$

上式により，文書中の単語におけるランキングを行うことができる．

実際に，トピックとして [Female]，[Male]，[Biological Markers] が付加されていたアブストラクトを例に以下説明する．このアブストラクトにおいて，各トピックの混合比は，それぞれ 0.499，0.460，0.041 と推定された．表 2 は，このアブストラクト中で，上記のエントロピーの低い順に単語を並べたものである．括弧'( )' の数値は，TF-IDF の高い順にランキングした場合の文書中での順位である (TF-IDF= $tf \cdot \log(M/df)$: tf は文書中での単語の出現頻度，df は対象としている単語を含む文書数，M は文書の総数).

表 2 Word List of Document whose Topics are [Female], [Male] and [Biological Markers]

Ranking	上位 10 語	Ranking	下位 10 語
1(37)	biomarkers	67(69)	indicate
2(19)	Fusarium	68(57)	problem
3(20)	non-Gaussian	69(45)	use
4(21)	Stachybotrys	70(75)	%
5(7)	chrysogenum	71(59)	correlate
6(22)	Cladosporium	72(17)	population
7(3)	mould	73(15)	healthy
8(35)	Aspergillus	7433)	response
9(23)	dampness	75(56)	man
10(24)	1SD	76(64)	woman

表 2 において，上位 10 語のほうが下位 10 語に比べてより専門性の高い単語であると考えられる．ある単語のトピックに対するエントロピーが低いということは，それだけ特定のトピックに特化した単語であると言える．したがって，上位に専門性の高い単語が現れると考えられる．また，この単語のランキングは，付加されるトピックとその混合比に依存する．エントロピーの計算に用いる ϕ は，式 (42) からもわかるとおりトピックの比率を示す γ によって変化する．このため付加されるトピックごとに単語のランキングが変化するのである．さらに，混合比も関連していることから文書毎に単語のランキングが変化すると考えられる．実際に，表 2 と同じアブストラクトに対して，でたらめにトピック [Rats]，[Child]，[Incidence] を付加して単語のランキングを行った例を表 3 に示す．各トピックの推定混合比は，それぞれ 0.411，0.352，0.237 であった．表 3 からも分かるとおり，文書中での単語のランキングが変化している．TF-IDF における順位は変わらない．

このような単語のランキングの変化は，TF-IDF などの単語の頻度情報のみを利用する手法では行うことはできない．

本提案モデルでは，上記のように文書に付加された多重トピックという情報を元に単語をランキングすることができる．従来のキーワード抽出手法と組み合わせることで，付加されたトピックや文書に特有のキーワードの抽出が行える可能性がある．

6. おわりに

本稿では，多重トピック文書における確率的生成モデル:Parametric Dirichlet Mixture Model(PDMM) を提案した．PDMM

表 3 Word List of Document whose Topics are [Rats], [Child] and [Incidence]

Ranking	上位 10 語	Ranking	下位 10 語
1(69)	indicate	67(56)	man
2(63)	relate	68(47)	blot
3(53)	antigen	69(6)	exposure
4(45)	use	70(54)	distribution
5(3)	mould	71(68)	evaluate
6(4)	versicolor	72(67)	examine
7(35)	Aspergillus	73(59)	correlate
8(7)	chrysogenum	74(58)	positive
9(8)	chartarum	75(1)	IgG
10(9)	herbarum	76(60)	adult

は，各トピックにおけるモデルパラメータを，Dirichlet 分布を事前確率と仮定した混合確率で混合し，階層ベイズモデルにより多重トピック文書をモデル化する．MEDLINE コーパスを使用し，多重トピック分類タスクによって，PDMM を評価したところ，PDMM の有効性を確認した．また，多重トピック情報を用いることで，混合比による文章の特徴量抽出や単語のランキング手法の可能性を示唆した．今後の予定としては，提案モデルと固有表現抽出などの従来手法との併用とともに上記の実際的な応用例を模索していく予定である．

謝辞 本研究は，文科省科学研究費 特定領域研究「情報爆発」の補助を得て行われた．

文　献

[1] D. M. Blei, Andrew Y. Ng, and M.I. Jordan."Latent Dirichlet Allocation," Neural Information Processing Systems 14, 2001.

[2] D. M. Blei, Andrew Y. Ng and M.I. Jordan. "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp.993-1022, 2003.

[3] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. "Hierarchical dirichlet processes," Technical Report 653, Department Of Statistics, UC Berkeley, 2003.

[4] Ueda, N. and Saito, K., "Parametric mixture models for multi-topic text," Neural Information Processing Systems 15(NIPS15), MIT Press, pp. 737-744, 2002.

[5] Ueda, N. and Saito, K., " Singleshot detection of multi-category text using parametric mixture models," ACM SIG Knowledge Discovery and Data Mining (SIGKDD2002), pp. 626-631, 2002.

[6] Minka ,"Estimating a Dirichlet distribution," Technical Report, 2002.
http://research.microsoft.com/ minka/papers/dirichlet/minka-dirichlet.pdf

[7] C.M.Bishop, "Pattern Recognition And Machine Learning (Information Science and Statistics," Springer-Verlag , p.687, 2006.

[8] H.Attias, "Learning parameters and structure of latent variable models by variational Bayes," in Proc of Uncertainty in Artificial Intelligence, 1999.

[9] 上田 修功, "ベイズ学習," 電子情報通信学会誌, vol.85, no.4,6,7,8 全 4 回, 2002

[10] 上田 修功, "計算統計 I(統計科学のフロンティア 11), " III 章 4 節 , 第 1 版, 2003