Bayesian Document Generative Model with Explicit Multiple Topics

Issei Sato
Graduate School of Information Science
and Technology,
The University of Tokyo
sato@r.dl.itc.u-tokyo.ac.jp

Hiroshi Nakagawa
Information Technology Center,
The University of Tokyo
nakagawa@dl.itc.u-tokyo.ac.jp

Abstract

In this paper, we proposed a novel probabilistic generative model to deal with explicit 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.

1 Introduction

Documents, such as those seen on Wikipedia and Folksonomy, have tended to be assigned with explicit multiple topics. In this situation, it is important to analyze a linguistic relationship between documents and the assigned multiple topics . We attempt to model this relationship with a probabilistic generative model. A probabilistic generative model for documents with multiple topics is a probability model of the process of generating documents with multiple topics. By focusing on modeling the generation process of documents and the assigned multiple topics, we can extract specific properties of documents and the assigned multiple topics. The model can also be applied to a wide range of applications such as automatic categorization for multiple topics, keyword extraction and measuring document similarity, for example.

A probabilistic generative model for documents with multiple topics is categorized into the following two models. One model assumes a topic as a latent topic. We call this model the latent-topic model. The other model assumes a topic as an explicit topic. We call this model the explicit-topic model.

In a latent-topic model, a latent topic indicates not a concrete topic but an underlying implicit topic of documents. Obviously this model uses an unsupervised learning algorithm. Representative examples of this kind of model are Latent Dirichlet Allocation(LDA)(D.M.Blei et al., 2001; D.M.Blei et al., 2003) and Hierarchical Dirichlet Process(HDP)(Y.W.Teh et al., 2003).

In an explicit-topic model, an explicit topic indicates a concrete topic such as economy or sports, for example. A learning algorithm for this model is a supervised learning algorithm. That is, an explicit topic model learns model parameter using a training data set of tuples such as (documents, topics). Representative examples of this model are Parametric Mixture Models(PMM1 and PMM2)(Ueda, N. and Saito, K., 2002a; Ueda, N. and Saito, K., 2002b). In the remainder of this paper, PMM indicates PMM1 because PMM1 is more effective than PMM2.

In this paper, we focus on the explicit topic model. In particular, we propose a novel model that is based on PMM but fundamentally improved.

The remaining part of this paper is organized as follows. Sections 2 explains terminology used in the

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 421–429, Prague, June 2007. ©2007 Association for Computational Linguistics

following sections. Section 3 explains PMM that is most directly related to our work. Section 4 points out the problem of PMM and introduces our new model. Section 5 evaluates our new model. Section 6 summarizes our work.

2 Terminology

This section explains terminology used in this paper. K is the number of explicit topics. V is the number of words in the vocabulary. $\mathscr{V} = \{1, 2, \cdots, V\}$ is a set of vocabulary index. $\mathscr{Y} = \{1, 2, \cdots, K\}$ is a set of topic index. N is the number of words in a document. $\boldsymbol{w} = (w_1, w_2, \cdots, w_N)$ is a sequence of N words where w_n denotes the nth word in the sequence. $\boldsymbol{w}$ is a document itself and is called words vector. $\boldsymbol{x} = (x_1, x_2, \cdots, x_V)$ is a word-frequency vector, that is, BOW(Bag Of Words) representation where x_v denotes the frequency of word v. w_n^v takes a value of 1(0) when w_n is $v \in \mathscr{V}$ (is not $v \in \mathscr{V}$). $\boldsymbol{y} = (y_1, y_2, \cdots, y_K)$ is a topic vector into which a document $\boldsymbol{w}$ is categorized, where y_i takes a value of 1(0) when the ith topic is (not) assigned with a document $\boldsymbol{w}$. $I_y \subset \mathscr{Y}$ is a set of topic index i, where y_i takes a value of 1 in $\boldsymbol{y}$. $\sum_{i \in I_y}$ and $\Pi_{i \in I_y}$ denote the sum and product for all i in I_y, respectively. $\Gamma(x)$ is the Gamma function and Ψ is the Psi function(Minka, 2002). A probabilistic generative model for documents with multiple topics models a probability of generating a document $\boldsymbol{w}$ in multiple topics $\boldsymbol{y}$ using model parameter Θ, i.e., models $P(\boldsymbol{w}|\boldsymbol{y}, \Theta)$. A multiple categorization problem is to estimate multiple topics $\boldsymbol{y}^*$ of a document $\boldsymbol{w}^*$ whose topics are unknown. The model parameters are learned by documents $D = \{(\boldsymbol{w_d}, \boldsymbol{y_d})\}_{d=1}^M$, where M is the number of documents.

3 Parametric Mixture Model

In this section, we briefly explain Parametric Mixture Model(PMM)(Ueda, N. and Saito, K., 2002a; Ueda, N. and Saito, K., 2002b).

3.1 Overview

PMM models multiple-topic documents by mixing model parameters of each single topic with an equal mixture ratio, where the model parameter θ_{iv} is the probability that word v is generated from topic i. This is because it is impractical to use model parameter corresponding to multiple topics whose number is $2^K - 1$(all combination of K topics). PMM achieved more useful results than machine learning methods such as Naive Bayes, SVM, K-NN and Neural Networks (Ueda, N. and Saito, K., 2002a; Ueda, N. and Saito, K., 2002b).

3.2 Formulation

PMM employs a BOW representation and is formulated as follows.

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

$\boldsymbol{\theta}$ is a $K \times V$ matrix whose element is $\theta_{iv} = P(v|y_i = 1)$. $\varphi(v, \boldsymbol{y}, \boldsymbol{\theta})$ is the probability that word v is generated from multiple topics $\boldsymbol{y}$ and is defined as the linear sum of $h_i(\boldsymbol{y})$ and θ_{iv} as follows: $\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^K h_i(\boldsymbol{y}) \theta_{iv}$

$h_i(\boldsymbol{y})$ is a mixture ratio corresponding to topic i and is formulated as follows:

$$h_i(\boldsymbol{y}) = \frac{y_i}{\sum_{j=1}^K y_j}, \sum_{i=1}^K h_i(\boldsymbol{y}) = 1.$$
$$\text{(if } y_i = 0\text{, then } h_i(\boldsymbol{y}) = 0)$$

3.3 Learning Algorithm of Model Parameter

The learning algorithm of model parameter $\boldsymbol{\theta}$ in PMM is an iteration method similar to the EM algorithm. Model parameter $\boldsymbol{\theta}$ is estimated by maximizing $\Pi_{d=1}^M P(\boldsymbol{w_d}|\boldsymbol{y_d}, \boldsymbol{\theta})$ in training documents $D = \{(\boldsymbol{w_d}, \boldsymbol{y_d})\}_{d=1}^M$. Function g corresponding to a document d is introduced as follows:

$$g_{iv}^d(\boldsymbol{\theta}) = \frac{h(\boldsymbol{y_d})\theta_{iv}}{\sum_{j=1}^K h_j(\boldsymbol{y_d})\theta_{jv}} \quad (2)$$

The parameters are updated along with the following formula.

$$\theta_{iv}^{(t+1)} = \frac{1}{C}(\sum_d^M x_{dv} g_{iv}^d(\boldsymbol{\theta}^{(t)}) + \zeta - 1) \quad (3)$$

x_{dv} is the frequency of word v in document d. C is the normalization term for $\sum_{v=1}^V \theta_{iv} = 1$. ζ is a smoothing parameter that is *Laplace smoothing* when ζ is set to two. In this paper, ζ is set to two as the original paper.

4 Proposed Model

In this section, firstly, we mention the problem related to PMM. Then, we explain our solution of the problem by proposing a new model.

4.1 Overview

PMM estimates model parameter $\boldsymbol{\theta}$ assuming that all of mixture ratios of single topic are equal. It is our intuition that each document can sometimes be more weighted to some topics than to the rest of the assigned topics. If the topic weightings are averaged over all biases in the whole document set, they could be canceled. Therefore, model parameter $\boldsymbol{\theta}$ learned by PMM can be reasonable over the whole of documents.

However, if we compute the probability of generating an individual document, a document-specific topic weight bias on mixture ratio is to be considered. The proposed model takes into account this document-specific bias by assuming that mixture ratio vector $\boldsymbol{\pi}$ follows Dirichlet distribution. This is because we assume the sum of the element in vector $\boldsymbol{\pi}$ is one and each element π_i is nonnegative. Namely, the proposed model assumes model parameter of multiple topics as a mixture of model parameter on each single topic with mixture ratio following Dirichlet distribution. Concretely, given a document $\boldsymbol{w}$ and multiple topics $\boldsymbol{y}$, it estimates a posterior probability distribution $P(\boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{y})$ by Bayesian inference. For convenience, the proposed model is called PDMM(Parametric Dirichlet Mixture Model).

In Figure 1, the mixture ratio(bias) $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3), \sum_{i=1}^{3} \pi_i = 1, \pi_i > 0$ of three topics is expressed in 3-dimensional real space $\boldsymbol{R}^3$. The mixture ratio(bias) $\boldsymbol{\pi}$ constructs 2D-simplex in $\boldsymbol{R}^3$. One point on the simplex indicates one mixture ratio $\boldsymbol{\pi}$ of the three topics. That is, the point indicates multiple topics with the mixture ratio. PMM generates documents assuming that each mixture ratio is equal. That is, PMM generates only documents with multiple topics that indicates the center point of the 2D-simplex in Figure 1. On the contrary, PDMM generates documents assuming that mixture ratio $\boldsymbol{\pi}$ follows Dirichlet distribution. That is, PDMM can generate documents with multiple topics whose weights can be generated by Dirichlet distribution.

4.2 Formulation

PDMM is formulated as follows:

$$P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$$

$$= \int P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y}) \Pi_{v=1}^{V} (\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi}))^{x_v} d\boldsymbol{\pi} \qquad (4)$$

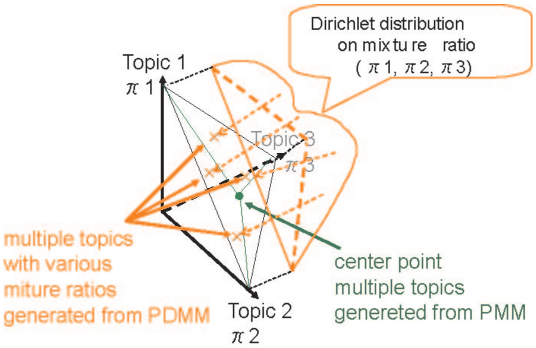

Figure 1: Topic Simplex for Three Topics

$\boldsymbol{\pi}$ is a vector whose element is $\pi_i (i \in I_y)$. π_i is a mixture ratio(bias) of model parameter corresponding to single topic i where $\pi_i > 0, \sum_{i \in I_y} \pi_i = 1$. π_i can be considered as a probability of topic i , i.e., $\pi_i = P(y_i = 1|\boldsymbol{\pi})$. $P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})$ is a prior distribution of $\boldsymbol{\pi}$ whose index i is an element of I_y, i.e., $i \in I_y$. We use Dirichlet distribution as the prior. $\boldsymbol{\alpha}$ is a parameter vector of Dirichlet distribution corresponding to $\pi_i (i \in I_y)$. Namely, the formulation is as follows.

$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y}) = \frac{\Gamma(\sum_{i \in I_y} \alpha_i)}{\Pi_{i \in I_y} \Gamma(\alpha_i)} \Pi_{i \in I_y} \pi_i^{\alpha_i - 1} \qquad (5)$$

$\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi})$ is the probability that word v is generated from multiple topics $\boldsymbol{y}$ and is denoted as a linear sum of $\pi_i (i \in I_y)$ and $\theta_{iv} (i \in I_y)$ as follows.

$$\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{i \in I_y} \pi_i \theta_{iv} \qquad (6)$$

$$= \sum_{i \in I_y} P(y_i = 1|\boldsymbol{\pi}) P(v|y_i = 1, \theta) \qquad (7)$$

4.3 Variational Bayes Method for Estimating Mixture Ratio

This section explains a method to estimate the posterior probability distribution $P(\boldsymbol{\pi}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ of a document-specific mixture ratio. Basically, $P(\boldsymbol{\pi}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is obtained by Bayes theorem using Eq.(4). However, that is computationally impractical because a complicated integral computation is needed. Therefore we estimate an approximate distribution of $P(\boldsymbol{\pi}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ using Variational Bayes Method(H.Attias, 1999). The concrete explanation is as follows

Use Eqs.(4)(7).

$$P(\boldsymbol{w}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) =$$
$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})\Pi_{v=1}^{V}(\sum_{i \in I_y} P(y_i = 1|\boldsymbol{\pi})P(v|y_i = 1, \theta))^{x_v}$$

Transform document expression of above equation into words vector $\boldsymbol{w} = (w_1, w_2, \cdots, w_N)$.

$$P(\boldsymbol{w}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) =$$
$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})\Pi_{n=1}^{N} \sum_{i_n \in I_y} P(y_{i_n} = 1|\boldsymbol{\pi})P(w_n|y_{i_n} = 1, \theta)$$

By changing the order of $\sum$ and Π, we have

$$P(\boldsymbol{w}, \boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) =$$
$$P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y}) \sum_{\boldsymbol{i} \in I_y^N} \Pi_{n=1}^{N} P(y_{i_n} = 1|\boldsymbol{\pi})P(w_n|y_{i_n} = 1, \theta)$$
$$(\sum_{\boldsymbol{i} \in I_y^N} \equiv \sum_{i_1 \in I_y} \sum_{i_2 \in I_y} \cdots \sum_{i_N \in I_y})$$

Express $y_{i_n} = 1$ as $z_n = i$.

$$P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) =$$
$$\int \sum_{\boldsymbol{z} \in I_y^N} P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})\Pi_{n=1}^{N} P(z_n|\boldsymbol{\pi})P(w_n|z_n, \theta)d\boldsymbol{\pi}$$
$$(\sum_{\boldsymbol{z} \in I_y^N} \equiv \sum_{z_1 \in I_y} \sum_{z_2 \in I_y} \cdots \sum_{z_N \in I_y}) \qquad (8)$$

Eq.(8) is regarded as Eq.(4) rewritten by introducing a new latent variable $\boldsymbol{z} = (z_1, z_2, \cdots, z_N)$.

$$P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \int \sum_{\boldsymbol{z} \in I_y^N} P(\boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})d\boldsymbol{\pi} \quad (9)$$

Use Eqs.(8)(9)

$$P(\boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$$
$$= P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})\Pi_{n=1}^{N} P(z_n|\boldsymbol{\pi})P(w_n|z_n, \theta) \qquad (10)$$

Hereafter, we explain Variational Bayes Method for estimating an approximate distribution of $P(\boldsymbol{\pi}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ using Eq.(10). This approach is the same as LDA(D.M.Blei et al., 2001; D.M.Blei et al., 2003). The approximate distribution is assumed to be $Q(\boldsymbol{\pi}, \boldsymbol{z}|\gamma, \boldsymbol{\phi})$. The following assumptions are introduced.

$$Q(\boldsymbol{\pi}, \boldsymbol{z}|\gamma, \boldsymbol{\phi}) = Q(\boldsymbol{\pi}|\gamma)Q(\boldsymbol{z}|\boldsymbol{\phi}) \qquad (11)$$
$$Q(\boldsymbol{\pi}|\gamma) = \frac{\Gamma(\sum_{i \in I_y} \gamma_i)}{\Pi_{i \in I_y}\Gamma(\gamma_i)} \Pi_{i \in I_y} \pi_i^{\gamma_i - 1} \qquad (12)$$
$$Q(\boldsymbol{z}|\boldsymbol{\phi}) = \Pi_{n=1}^{N} Q(z_n|\boldsymbol{\phi}) \qquad (13)$$
$$Q(z_n|\boldsymbol{\phi}) = \Pi_{i=1}^{K} (\phi_{ni})^{z_n^i} \qquad (14)$$

$Q(\boldsymbol{\pi}|\gamma)$ is Dirichlet distribution where γ is its parameter. $Q(z_n|\boldsymbol{\phi})$ is Multinomial distribution where ϕ_{ni} is its parameter and indicates the probability that the nth word of a document is topic i, i.e. $P(y_{i_n} = 1)$. z_n^i is a value of 1(0) when z_n is (not) i. According to Eq.(11), $Q(\boldsymbol{\pi}|\gamma)$ is regarded as an approximate distribution of $P(\boldsymbol{\pi}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$

The log likelihood of $P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is derived as follows.

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

$$\log P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \mathscr{F}[Q] + KL(Q, P) \qquad (15)$$

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

$KL(Q, P)$ is the Kullback-Leibler Divergence that is often employed as a distance between probability distributions. Namely, $KL(Q, P)$ indicates a distance between $Q(\boldsymbol{\pi}, \boldsymbol{z}|\gamma, \boldsymbol{\phi})$ and $P(\boldsymbol{\pi}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. $\log P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is not relevant to $Q(\boldsymbol{\pi}, \boldsymbol{z}|\gamma, \boldsymbol{\phi})$. Therefore, $Q(\boldsymbol{\pi}, \boldsymbol{z}|\gamma, \boldsymbol{\phi})$ that maximizes $\mathscr{F}[Q]$ minimizes $KL(Q, P)$, and gives a good approximate distribution of $P(\boldsymbol{\pi}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\theta})$.

We estimate $Q(\boldsymbol{\pi}, \boldsymbol{z}|\gamma, \boldsymbol{\phi})$, concretely its parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, by maximizing $\mathscr{F}[Q]$ as follows.

Using Eqs.(10)(11).

$$\mathscr{F}[Q] = \int Q(\boldsymbol{\pi}|\gamma) \log P(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{y})d\boldsymbol{\theta} \qquad (16)$$
$$+ \int \sum_{\boldsymbol{z} \in I_y^N} Q(\boldsymbol{\pi}|\gamma)Q(\boldsymbol{z}|\boldsymbol{\phi}) \log \Pi_{n=1}^{N} P(z_n|\boldsymbol{\pi})d\boldsymbol{\theta} \quad (17)$$
$$+ \sum_{\boldsymbol{z} \in I_y^N} Q(\boldsymbol{z}|\boldsymbol{\phi}) \log \Pi_{n=1}^{N} P(w_n|z_n, \theta) \qquad (18)$$
$$- \int Q(\boldsymbol{\pi}|\gamma) \log Q(\boldsymbol{\pi}|\gamma)d\boldsymbol{\theta} \qquad (19)$$
$$- \sum_{\boldsymbol{z} \in I_y^N} Q(\boldsymbol{z}|\boldsymbol{\phi}) \log Q(\boldsymbol{z}|\boldsymbol{\phi}) \qquad (20)$$

$$= \quad \log\Gamma(\textstyle\sum_{i\in I_y}\alpha_j) - \sum_{i\in I_y}\log\Gamma(\alpha_i) + \sum_{i\in I_y}(\alpha_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j))(21)$$

$$+ \quad \sum_{n=1}^{N}\sum_{i\in I_y}\phi_{ni}(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) \quad (22)$$

$$+ \quad \sum_{n=1}^{N}\sum_{i\in I_y}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\theta ij \quad (23)$$

$$- \quad \log\Gamma(\sum_{j\in I_y}\gamma_j) + \sum_{i\in I_y}\log\Gamma(\sum_{j\in I_y}\gamma_j)$$

$$-\sum_{i\in I_y}(\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) \, (24)$$

$$- \quad \sum_{n=1}^{N}\sum_{i\in I_y}\phi_{ni}\log\phi_{ni} \quad (25)$$

$\mathscr{F}[Q]$ is known to be a function of γ_i and ϕ_{ni} from Eqs.(21) through (25). Then we only need to resolve the maximization problem of nonlinear function $\mathscr{F}[Q]$ with respect to γ_i and ϕ_{ni}. In this case, the maximization problem can be resolved by Lagrange multiplier method.

First, regard $\mathscr{F}[Q]$ as a function of γ_i, which is denoted as $\mathscr{F}[\gamma_i]$. Then ,γ_i does not have constraints. Therefore we only need to find the following γ_i, where $\frac{\partial\mathscr{F}[\gamma_i]}{\partial\gamma_i} = 0$. The resultant γ_i is expressed as follows.

$$\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni} \quad (i \in I_y) \quad (26)$$

Second, regard $\mathscr{F}[Q]$ as a function of ϕ_{ni}, which is denoted as $\mathscr{F}[\phi_{ni}]$. Then, considering the constraint that $\sum_{i\in I_y}\phi_{ni} = 1$, Lagrange function $L[\phi_{ni}]$ is expressed as follows:

$$L[\phi_{ni}] = \mathscr{F}[\phi_{ni}] + \lambda(\sum_{i\in I_y}\phi_{ni} - 1) \quad (27)$$

λ is a so-called Lagrange multiplier.

We find the following ϕ_{ni} where $\frac{\partial L[\phi_{ni}]}{\partial\phi ni} = 0$.

$$\phi_{ni} = \frac{\theta_{iw_n}}{C}\exp(\Psi(\gamma_i) - \Psi(\sum_{j\in I_y}\gamma_j)) \;\; (i \in I_y))\,(28)$$

C is a normalization term. By Eqs.(26)(28), we obtain the following updating formulas of γ_i and ϕ_{ni}.

$$\gamma_i^{(t+1)} = \alpha_i + \sum_{n=1}^{N}\phi_{ni}^{(t)} \quad (i \in I_y) \quad (29)$$

$$\phi_{ni}^{(t+1)} = \frac{\theta_{iw_n}}{C}\exp(\Psi(\gamma_i^{(t+1)}) - \Psi(\sum_{j\in I_y}\gamma_j^{(t+1)}))\,(30)$$

Using the above updating formulas , we can estimate parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, which are specific to a document $\boldsymbol{w}$ and topics $\boldsymbol{y}$. Last of all , we show a pseudo code :$vb(\boldsymbol{w},\boldsymbol{y})$ which estimates γ and ϕ. In addition , we regard $\boldsymbol{\alpha}$, which is a parameter of a prior distribution of $\boldsymbol{\pi}$, as a vector whose elements are all one. That is because Dirichlet distribution where each parameter is one becomes Uniform distribution.

```
• Variational Bayes Method for PDMM————
function vb($\boldsymbol{w},\boldsymbol{y}$):

1. Initialize $\alpha_i \leftarrow 1 \; \forall i \in I_y$
2. Compute $\boldsymbol{\gamma^{(t+1)}}, \boldsymbol{\phi^{(t+1)}}$ using Eq.(29)(30)
3. if $\parallel \boldsymbol{\gamma^{(t+1)}} - \boldsymbol{\gamma^{(t)}} \parallel < \epsilon$
   & $\parallel \boldsymbol{\phi^{(t+1)}} - \boldsymbol{\phi^{(t)}} \parallel < \epsilon$
4. then return $(\boldsymbol{\gamma^{(t+1)}}, \boldsymbol{\phi^{(t+1)}})$ and halt
5. else $t \leftarrow t + 1$ and goto step (2)
```

4.4 Computing Probability of Generating Document

PMM computes a probability of generating a document $\boldsymbol{w}$ on topics $\boldsymbol{y}$ and a set of model parameter Θ as follows:

$$P(\boldsymbol{w}|\boldsymbol{y},\Theta) = \Pi_{v=1}^{V}(\varphi(v,\boldsymbol{y},\boldsymbol{\theta}))^{x_v} \quad (31)$$

$\varphi(v,\boldsymbol{y},\boldsymbol{\theta})$ is the probability of generating a word v on topics $\boldsymbol{y}$ that is a mixture of model parameter $\theta_{iv}(i \in I_y)$ with an equal mixture ratio. On the other hand, PDMM computes the probability of generating a word v on topics $\boldsymbol{y}$ using $\theta_{iv}(i \in I_y)$ and an approximate posterior distribution $Q(\pi|\boldsymbol{\gamma})$ as follows:

$$\begin{aligned}
&\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\
&\quad = \int \Big(\sum_{i \in I_y} \pi_i \theta_{iv}\Big) Q(\boldsymbol{\pi}|\boldsymbol{\gamma}) d\boldsymbol{\pi} && (32) \\
&\quad = \sum_{i \in I_y} \int \pi_i Q(\pi|\boldsymbol{\gamma}) d\boldsymbol{\pi} \theta_{iv} && (33) \\
&\quad = \sum_{i \in I_y} \tilde{\pi}_i \theta_{iv} && (34)
\end{aligned}$$

$\tilde{\pi}_i = \int \pi_i Q(\pi|\boldsymbol{\gamma}) d\boldsymbol{\pi} = \frac{\gamma_i}{\sum_{j \in I_y} \gamma_j}$ (C.M.Bishop, 2006)

The above equation regards the mixture ratio of topics $\boldsymbol{y}$ of a document $\boldsymbol{w}$ as the expectation $\tilde{\pi}_i (i \in I_y)$ of $Q(\pi|\boldsymbol{\gamma})$. Therefore, a probability of generating $\boldsymbol{w}$ $P(\boldsymbol{w}|\boldsymbol{y}, \Theta)$ is computed with $\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ estimated in the following manner:

$$P(\boldsymbol{w}|\boldsymbol{y}, \Theta) = \Pi_{v=1}^{V} (\varphi(v, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\gamma})))^{x_v} \qquad (35)$$

4.5 Algorithm for Estimating Multiple Topics of Document

PDMM estimates multiple topics $\boldsymbol{y}^*$ maximizing a probability of generating a document $\boldsymbol{w}^*$, i.e., Eq.(35). This is the 0-1 integer problem(i.e., NP-hard problem), so PDMM uses the same approximate estimation algorithm as PMM does. But it is different from PMM's estimation algorithm in that it estimates the mixture ratios of topics $\boldsymbol{y}$ by Variational Bayes Method as shown by vb(w,y) at step 6 in the following pseudo code of the estimation algorithm:

• Topics Estimation 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
```

5 Evaluation

We evaluate the proposed model by using F-measure of multiple topics categorization problem.

5.1 Dataset

We use MEDLINE[1] as a dataset. In this experiment, we use five thousand abstracts written in English. MEDLINE has a metadata set called MeSH Term. For example, each abstract has MeSH Terms such as RNA Messenger and DNA-Binding Proteins. MeSH Terms are regarded as multiple topics of an abstract. In this regard, however, we use MeSH Terms whose frequency are medium(100-999). We did that because the result of experiment can be overly affected by such high frequency terms that appear in almost every abstract and such low frequency terms that appear in very few abstracts. In consequence, the number of topics is 88. The size of vocabulary is 46,075. The proportion of documents with multiple topics on the whole dataset is 69.8%, i.e., that of documents with single topic is 30.2%. The average of the number of topics of a document is 3.4. Using TreeTagger[2], we lemmatize every word. We eliminate stop words such as articles and be-verbs.

5.2 Result of Experiment

We compare F-measure of PDMM with that of PMM and other models.

F-measure(F) is as follows:
$F = \frac{2PR}{P+R}, P = \frac{|N_r \cap N_e|}{|N_e|}, R = \frac{|N_r \cap N_e|}{|N_r|}$.

N_r is a set of relevant topics . N_e is a set of estimated topics. A higher F-measure indicates a better ability to discriminate topics. In our experiment, we compute F-measure in each document and average the F-measures throughout the whole document set.

We consider some models that are distinct in learning model parameter $\boldsymbol{\theta}$. **PDMM** learns model parameter $\boldsymbol{\theta}$ by the same learning algorithm as PMM. **NBM** learns model parameter $\boldsymbol{\theta}$ by Naive Bayes learning algorithm. The parameters are updated according to the following formula: $\theta_{iv} = \frac{M_{iv}+1}{C}$. M_{iv} is the number of training documents where a word v appears in topic i. C is normalization term for $\sum_{v=1}^{V} \theta_{iv} = 1$.

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

The comparison of these models with respect to F-measure is shown in Figure 2. The horizontal axis is the proportion of test data of dataset(5,000 abstracts). For example, 2% indicates that the number of documents for learning model is 4,900 and the number of documents for the test is 100. The vertical axis is F-measure. In each proportion, F-measure is an average value computed from five pairs of training documents and test documents randomly generated from dataset.

F-measure of PDMM is higher than that of other methods on any proportion, as shown in Figure 2. Therefore, PDMM is more effective than other methods on multiple topics categorization.

Figure 3 shows the comparison of models with respect to F-measure, changing proportion of multiple topic document for the whole dataset. The proportion of document for learning and test are 40% and 60%, respectively. The horizontal axis is the proportion of multiple topic document on the whole dataset. For example, 30% indicates that the proportion of multiple topic document is 30% on the whole dataset and the remaining documents are single topic , that is, this dataset is almost single topic document. In 30%. there is little difference of F-measure among models. As the proportion of multiple topic and single topic document approaches 90%, that is, multiple topic document, the differences of F-measure among models become apparent. This result shows that PDMM is effective in modeling multiple topic document.

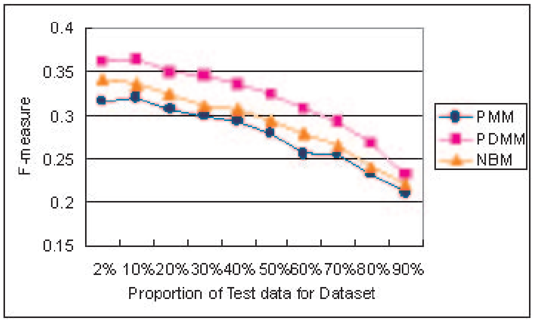

Figure 2: F-measure Results

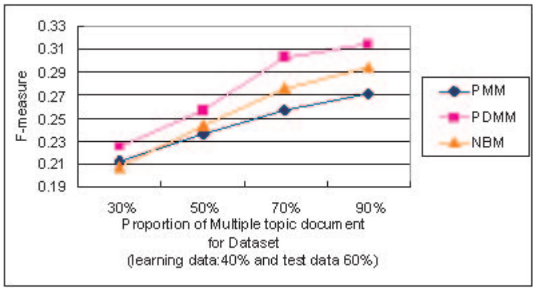

Figure 3: F-measure Results changing Proportion of Multiple Topic Document for Dataset

5.3 Discussion

In the results of experiment described in section 5.2, PDMM is more effective than other models in multiple-topic categorization. If the topic weightings are averaged over all biases in the whole of training documents, they could be canceled. This cancellation can lead to the result that model parameter $\boldsymbol{\theta}$ learned by PMM is reasonable over the whole of documents. Moreover, PDMM computes the probability of generating a document using a mixture of model parameter, estimating the mixture ratio of topics. This estimation of the mixture ratios, we think, is the key factor to achieve the results better than other models. In addition, the estimation of a mixture ratio of topics can be effective from the perspective of extracting features of a document with multiple topics. A mixture ratio of topics assigned to a document is specific to the document. Therefore, the estimation of the mixture ratio of topics is regarded as a projection from a word-frequency space of $\mathscr{Q}^V$ where $\mathscr{Q}$ is a set of integer number to a mixture ratio space of topics $[0,1]^K$ in a document. Since the size of vocabulary is much more than that of topics, the estimation of the mixture ratio of topics is regarded as a dimension reduction and an extraction of features in a document. This can lead to analysis of similarity among documents with multiple topics. For example, the estimated mixture ratio of topics [Comparative Study]C[Apoptosis] and [Models,Biological] in one MEDLINE abstract is 0.656C0.176 and 0.168, respectively. This ratio can be a feature of this document.

Moreover, we can obtain another interesting results as follows. The estimation of mixture ratios of topics uses parameter $\boldsymbol{\gamma}$ in section 4.3. We obtain interesting results from another parameter $\boldsymbol{\phi}$ that needs to estimate $\boldsymbol{\gamma}$. ϕ_{ni} is specific to a document. A

Table 1: Word List of Document X whose Topics are [Female], [Male] and [Biological Markers]

Ranking	Top10	Ranking	Bottom10
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

Table 2: Word List of Document X whose Topics are [Rats], [Child] and [Incidence]

Ranking	Top10	Ranking	Bottom10
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

ϕ_{ni} indicates the probability that a word w_n belongs to topic i in a document. Therefore we can compute the entropy on w_n as follows:

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

We rank words in a document by this entropy. For example, a list of words in ascending order of the entropy in document X is shown in Table 1. A value in parentheses is a ranking of words in decending order of TF-IDF($= tf \cdot \log(M/df)$,where tf is term frequency in a test document, df is document frequency and M is the number of documents in the set of doucuments for learning model parameters) (Y. Yang and J. Pederson, 1997) . The actually assigned topics are [Female] , [Male] and [Biological Markers], where each estimated mixture ratio is 0.499 , 0.460 and 0.041, respectively.

The top 10 words seem to be more technical than the bottom 10 words in Table 1. When the entropy of a word is lower, the word is more topic-specific oriented, i.e., more technical . In addition, this ranking of words depends on topics assigned to a document. When we assign randomly chosen topics to the same document, generic terms might be ranked higher. For example, when we rondomly assign the topics [Rats], [Child] and [Incidence], generic terms such as ”use” and ”relate” are ranked higher as shown in Table 2. The estimated mixture ratio of [Rats], [Child] and [Incidence] is 0.411, 0.352 and 0.237, respectively.

For another example, a list of words in ascending order of the entropy in document Y is shown in Table 3. The actually assigned topics are Female, Animals, Pregnancy and Glucose.. The estimated mixture ratio of [Female], [Animals] ,[Pregnancy] and [Glucose] is 0.442, 0.437, 0.066 and 0.055, respectively In this case, we consider assigning sub topics of actual topics to the same document Y.

Table 4 shows a list of words in document Y assigned with the sub topics [Female] and [Animals]. The estimated mixture ratio of [Female] and [Animals] is 0.495 and 0.505, respectively. Estimated mixture ratio of topics is chaged. It is interesting that [Female] has higher mixture ratio than [Animals] in actual topics but [Female] has lower mixture ratio than [Animals] in sub topics [Female] and [Animals]. According to these different mixture ratios, the ranking of words in docment Y is changed.

Table 5 shows a list of words in document Y assigned with the sub topics [Pregnancy] and [Glucose]. The estimated mixture ratio of [Pregnancy] and [Glucose] is 0.502 and 0.498, respectively. It is interesting that in actual topics, the ranking of gglucose-insulinh and ”IVGTT” is high in document Y but in the two subset of actual topics, gglucose-insulinh and ”IVGTT” cannot be find in Top 10 words.

The important observation known from these examples is that this ranking method of words in a document can be assosiated with topics assigned to the document. ϕ depends on γ seeing Eq.(28). This is because the ranking of words depends on assigned topics, concretely, mixture ratios of assigned topics. TF-IDF computed from the whole documents cannot have this property. Combined with existing the extraction method of keywords, our model has the potential to extract document-specific keywords using information of assigned topics.

Table 3: Word List of Document Y whose Actual Topics are [Femaile],[Animals],[Pregnancy] and [Glucose]

Ranking	Top 10	Ranking	Bottom 10
1(2)	glucose-insulin	94(93)	assess
2(17)	IVGTT	95(94)	indicate
3(11)	undernutrition	96(74)	CT
4(12)	NR	97(28)	%
5(13)	NRL	98(27)	muscle
6(14)	GLUT4	99(85)	receive
7(56)	pregnant	100(80)	status
8(20)	offspring	101(100)	protein
9(31)	pasture	102(41)	age
10(32)	singleton	103(103)	conclusion

Table 4: Word List of Document Y whose Topics are [Femaile]and [Animals]

Ranking	Top 10	Ranking	Bottom 10
1(31)	pasture	94(65)	insulin
2(32)	singleton	95(76)	reduced
3(33)	insulin-signaling	96(27)	muscle
4(34)	CS	97(74)	CT
5(35)	euthanasia	98(68)	feed
6(36)	humane	99(100)	protein
7(37)	NRE	100(80)	status
8(38)	110-term	101(85)	receive
9(50)	insert	102(41)	age
10(11)	undernutrition	103(103)	conclusion

6 Concluding Remarks

We proposed and evaluated a novel probabilistic generative models, PDMM, to deal with multiple-topic documents. We evaluated PDMM and other models by comparing F-measure using MEDLINE corpus. The results showed that PDMM is more effective than PMM. Moreover, we indicate the potential of the proposed model that extracts document-specific keywords using information of assigned topics.

Acknowledgement This research was funded in part by MEXT Grant-in-Aid for Scientific Research on Priority Areas ”i-explosion” in Japan.

Table 5: Word List of Document Y whose Topics are [Pregnancy]and [Glucose]

Ranking	Top 10	Ranking	Bottom 10
1(84)	mass	94(18)	IVGTT
2(74)	CT	95(72)	metabolism
3(26)	requirement	96(73)	metabolic
4(45)	intermediary	97(57)	pregnant
5(50)	insert	98(58)	prenatal
6(53)	feeding	99(59)	fetal
7(55)	nutrition	100(3)	gestation
8(61)	nutrient	101(20)	offspring
9(31)	pasture	102(65)	insulin
10(32)	singleton	103(16)	glucose

References

H.Attias 1999. Learning parameters and structure of latent variable models by variational Bayes. *in Proc of Uncertainty in Artificial Intelligence*.

C.M.Bishop 2006. Pattern Recognition And Machine Learning (Information Science and Statistics), p.687. *Springer-Verlag*.

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

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

Minka 2002. Estimating a Dirichlet distribution. *Technical Report*.

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

Ueda, N. and Saito, K. 2002. Parametric mixture models for multi-topic text. *Neural Information Processing Systems* 15.

Ueda, N. and Saito, K. 2002. Singleshot detection of multi-category text using parametric mixture models. *ACM SIG Knowledge Discovery and Data Mining*.

Y. Yang and J. Pederson 1997. A comparative study on feature selection in text categorization. *Proc. International Conference on Machine Learning*.