Dirichlet Process Mixture Model for Document Clustering with Feature Partition

Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi

Abstract—Finding the appropriate number of clusters to which documents should be partitioned is crucial in document clustering. In this paper, we propose a novel approach, namely DPMFP, to discover the latent cluster structure based on the DPM model without requiring the number of clusters as input. Document features are automatically partitioned into two groups, in particular, discriminative words and nondiscriminative words, and contribute differently to document clustering. A variational inference algorithm is investigated to infer the document collection structure as well as the partition of document words at the same time. Our experiments indicate that our proposed approach performs well on the synthetic data set as well as real data sets. The comparison between our approach and state-of-the-art document clustering approaches shows that our approach is robust and effective for document clustering.

Index Terms—Database management, database applications-text mining, pattern recognition, clustering document clustering, Dirichlet process mixture model, feature partition

1 Introduction

Document clustering, grouping unlabeled text documents into meaningful clusters, is of substantial interest in many applications. One assumption, taken by traditional document clustering approaches, as in [1], [2], [3], is that the number of clusters K is known before the process of document clustering. K is regarded as a predefined parameter determined by users. However, in reality, determining the appropriate value of K is a difficult problem. First, given a set of documents, users have to browse the whole document collection in order to estimate K. This is not only time consuming but also unrealistic especially when dealing with large document data sets. Furthermore, an improper estimation of K might easily mislead the clustering process. Clustering accuracy degrades drastically if a bigger or a smaller number of clusters is used. Therefore, it is very useful if a document clustering approach could be designed relaxing the assumption of the predefined K.

In this paper, we attempt to group documents into an optimal number of clusters while the number of clusters K is discovered automatically. The first contribution of our approach is to develop a Dirichlet Process Mixture (DPM) model to partition documents. The DPM model has been studied in nonparametric Bayesian for a long time [4], [5], [6]. It shows promising results for the clustering problem when the number of clusters is unknown. The basic idea of DPM model is to jointly consider both the data likelihood and the clustering property of the Dirichlet Process (DP) prior that data points are more likely to be related to popular and large clusters [7], [8]. When a new data point arrives, it either rises from existing cluster or starts a new cluster. This flexibility of the DPM model makes it particularly promising for document clustering. However, in the literature, there is little work investigating DPM model for document clustering due to the high-dimensional representation of text documents. In the problem of document clustering, each document is represented by a large amount of words including discriminative words and nondiscriminative words. Only discriminative words are useful for grouping documents. The involvement of nondiscriminative words confuses the clustering process and leads to poor clustering solution in return [9]. When the number of clusters is unknown, the affect of nondiscriminative words is aggravated.

The second contribution of our approach is to address this issue and design a DPM model to tackle the problem of document clustering. A novel model, namely DPMFP, is investigated which extends the traditional DPM model by conducting feature partition. Words in documents set are partitioned into two groups, in particular, discriminative words and nondiscriminative words. Each document is regarded as a mixture of two components. The first component, discriminative words, are generated from the specific cluster to which document belongs. The second component, nondiscriminative words, are generated from a general background shared by all documents. Only discriminative words are used to infer the latent cluster structure.

The computational cost of DPM parameter estimation is also a problem for developing the DPM model for the document clustering problem. Traditionally, there are two algorithms to infer DPM parameters, in particular, the

- *R. Huang is with the College of Computer Science and Information, Guizhou University, Guiyang 55000, China, and the Department of Industrial and System Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong. E-mail: cse,rzhuang@gzu.edu.cn.*
- *G. Yu is with the Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, NC 27599. E-mail: guanyu@email.unc.edu, yuguan1987@mail.nankai.edu.cn.*
- *Z. Wang is with the Department of Statistics, Nankai University, Tianjin 300071, China. E-mail: zjwang@nankai.edu.cn.*
- *J. Zhang is with the Department of Computer Science, Sun Yat-sen University, Guangzhou 510275, China. E-mail: junzhang@ieee.org.*
- *L. Shi is with the College of Management and Economics, Tianjin University, Tianjin 300072, China. E-mail: shi@tju.edu.cn.*

Manuscript received 25 Feb. 2011; revised 26 Oct. 2011; accepted 22 Jan. 2012; published online 13 Feb. 2012.
Recommended for acceptance by J. Zobel.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2011-02-0085.
Digital Object Identifier no. 10.1109/TKDE.2012.27.

variational inference algorithm and the Gibbs sampling algorithm. It is hard to apply the Gibbs sampling algorithm to document clustering since it needs long time to converge. Due to the high-dimensional representation of text documents, it is even harder to be applied when the document data set is large. For the algorithm of variational inference, it could be applied to infer the document collection structure in a much quicker manner. However, in our DPMFP approach, we need to infer the document collection structure as well as the the partition of document words at the same time. Therefore, traditional variation inference algorithm for the DPM model cannot be directly applied to our problem. The third contribution of our approach is to design a method to estimate the document collection structure for the DPMFP model. A Dirichlet Multinomial Allocation (DMA) model, namely DMAFP, is used to approximate the DPMFP model to simplify the process of parameter estimation. A variational inference algorithm is then derived for the DMAFP model. The Gibbs sampling algorithm is also investigated for comparison.

We have conducted extensive experiments on our proposed approach by using both synthetic and realistic data sets. We also compared our approach with state-of-the-art model-based clustering algorithms [1], [3]. Experimental results show that our proposed approach is robust and effective for document clustering.

The remainder of this paper is organized as follows: Section 2 reviews the related work on the document clustering and the identification of the number of clusters. In Section 3, we introduce the background knowledge of the DPM model. In Section 4, we describe our proposed DPMFP model and its DMAFP approximation. Our proposed variational inference algorithm and the Gibbs sampling algorithm are given in Section 5. Section 6 presents the experimental results. We finally present conclusions and future work in Section 7.

2 Related Work

Document clustering methods can be categorized based on whether the number of clusters is required as the input parameter. If the number of clusters is predefined, many algorithms based on the probabilistic finite mixture model have been provided in the literature. Nigam et al. [3] proposed a multinomial mixture model. It applies the EM algorithm for document clustering assuming that document topics follow multinomial distribution. Deterministic annealing procedures [10] are proposed to allow this algorithm to find better local optima of the likelihood function. Though multinomial distribution is often used to model text document, it fails to account for the burstiness phenomenon that if a word occurs once in a document, it is likely to occur repeatedly. Madsen et al. [2] used the DCM model to capture burstiness well. Its experiments showed that the performance of DCM was comparable to that obtained with multiple heuristic changes to the multinomial model. However, DCM model lacks intuitiveness and the parameters in that model cannot be estimated quickly. Elkan [1] derived the EDCM distribution which belongs to the exponential family. It is a good approximation to the DCM distribution. The EM algorithm with the EDCM distributions is much faster than the corresponding algorithm with DCM distributions proposed in [2]. It also attains high clustering accuracy. In recent years, EM algorithm with EDCM distribution is the most competitive algorithm for document clustering if the number of clusters is predefined.

If the number of clusters K is unknown before the clustering process, one solution is to estimate K first and use this estimation as the input parameter for those document clustering algorithms requiring K predefined. Many methods have been introduced to find an estimation of K. The most straightforward method is the likelihood cross-validation technique [11], which trains the model with different values of K and picks the one with the highest likelihood on some held-out data. Another method is to assign a prior to K and then calculate the posterior distribution of K to determine its value [12]. In the literature, there are also many information criteria proposed to choose K, e.g., Minimum Description Length (MDL) [13], Minimum Message Length (MML) [14], Akaike Information Criterion (AIC) [15], and Bayesian Information Criteria (BIC) [16]. The basic idea of all these criteria is to penalize complicated models (i.e., models with large K) in order to come up with an appropriate K to tradeoff data likelihood and model complexity [17].

An alternative solution is to use the DPM model which infers the number of clusters and the latent clustering structure simultaneously. The number of clusters is determined in the clustering process rather than preestimated. In our preliminary work, we proposed the DPMFS approach [18] using the DPM model to model the documents. A Gibbs Sampling algorithm was provided to infer the cluster structure. However, as the other MCMC methods, the Gibbs sampling method for the DPMFS model is slow to converge and its convergence is difficult to diagnose. Furthermore, it's difficult for us to develop effective variational inference method for the DPMFS model. Our proposed new model and the associated variational inference method in this paper solves these problems successfully.

3 Background

3.1 Dirichlet Process Mixture Model

The DPM model is a flexible mixture model in which the number of mixture components grows as new data are observed. It is one kind of countably infinite mixture model [19]. We introduce this infinite mixture model by first describing the simple finite mixture model.

In the finite mixture model, each data point is drawn from one of K fixed unknown distributions. For example, the multinomial mixture model for document clustering assumes that each document x_d is drawn from one of K multinomial distributions. Let η_d be the parameter of the distribution from which the document x_d is generated. Since the number of clusters is always unknown, to allow it to grow with data, we assume that the data point x_d follows a general mixture model in which η_d is generated from a distribution G. The conditional hierarchical relationships are as follows:

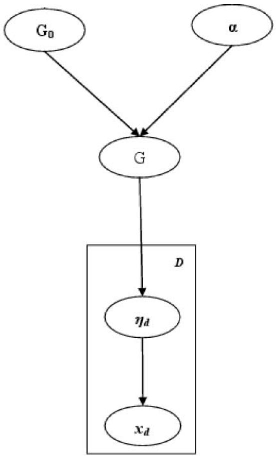

Fig. 1. Graphical representation of the DPM model.

$$\eta_d|G \sim G, d = 1, 2, \ldots, D,$$
$$x_d|\eta_d \sim F(x_d|\eta_d), d = 1, 2, \ldots, D,$$

where D is the number of data points and $F(x_d|\eta_d)$ is the distribution of x_d given η_d.

The probability distribution G mentioned above is always unknown. If G is a discrete distribution on a finite set of values, this generative mixture model reduces to the finite mixture model. In the nonparametric Bayesian analysis, the Dirichlet process mixture model places a Dirichlet process prior on the unknown distribution G. In this way, G can be considered as a mixture distribution with a random number of components. More formally, the hierarchical Bayesian specification of the DPM model is as follows:

$$G|\alpha, G_0 \sim DP(\alpha, G_0),$$
$$\eta_d|G \sim G, d = 1, 2, \ldots, D,$$
$$x_d|\eta_d \sim F(x_d|\eta_d), d = 1, 2, \ldots, D,$$

where $DP(\alpha, G_0)$ represents a DP with a base distribution G_0 and a positive scaling parameter α. Intuitively, G_0 is the mean of the DP and α is the inverse variance. G is more similar with G_0 when a larger value is assigned to α.

Since G is viewed as a random probability distribution in the DPM model, integrating out G, the joint distribution of the collection of variables $\eta_1, \eta_2, \ldots, \eta_D$ exhibits a clustering effect [4]. Let η_{-d} denote the set of η_j for $j \neq d$. Conditioning on η_{-d}, the distribution of η_d has the following form:

$$\eta_d|\eta_{-d}, \alpha, G_0 \sim \sum_{j\neq d} \frac{\delta_{\eta_j}}{D-1+\alpha} + \frac{1}{D-1+\alpha} G_0. \quad (1)$$

Let $\eta_1^*, \eta_2^*, \ldots, \eta_C^*$ denote the distinct values of $\eta_1, \eta_2, \ldots, \eta_D$. Let m_i be the number of times that η_i^* occurs in η_{-d}. The conditional distribution of η_d given η_{-d} follows the $Po^{'}lya$ urn distribution as follows:

$$\eta_d|\eta_{-d}, \alpha, G_0 \sim \sum_{i=1}^{C} \frac{m_i}{D-1+\alpha} \delta_{\eta_i^*} + \frac{1}{D-1+\alpha} G_0. \quad (2)$$

Equation (2) indicates that the data point x_d is either allocated to an existing cluster or a new cluster. In particular, x_d can be assigned to an existing cluster with the probability proportional to the cluster size or a new cluster with probability proportional to α. The number of clusters is determined automatically. We can best understand this clustering property by a famous metaphor known as the Chinese restaurant process [19]. The hierarchical representation of the DPM model is shown in Fig. 1.

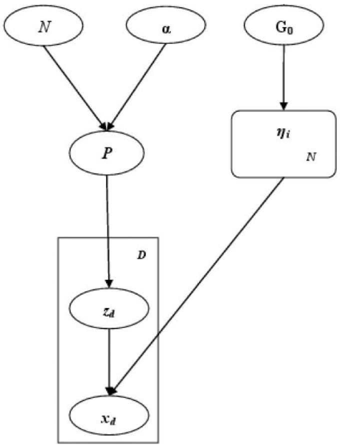

Fig. 2. Graphical representation of the DMA model.

3.2 Dirichlet Multinomial Allocation Model

It has been shown that the DPM model can be derived as the limit of a sequence of finite mixture models when the number of mixture components is taken to infinity [20], [21]. One famous approximation to the DPM model is the Dirichlet Multinomial Allocation (DMA) model [20]. The generative model for the DMA model is as follows:

$$P \sim Dirichlet(\alpha/N, \ldots, \alpha/N),$$
$$\eta_i \sim G_0, i = 1, 2, \ldots, N,$$
$$z_d|P \sim Discrete(p_1, p_2, \ldots, p_N), d = 1, 2, \ldots, D,$$
$$x_d|z_d, \eta_1, \eta_2, \ldots, \eta_N \sim F(x_d|\eta_{z_d}), d = 1, 2, \ldots, D,$$

where N is the number of mixture components. P is a N-dimensional vector indicating the mixing proportions for components given a Dirichlet prior with symmetric parameters α/N. z_d is an integer indicating the latent component allocation of the data point x_d. For each component, the parameter η_i determines the distribution of data points from that component. The graphical representation of the DMA model is shown in Fig. 2.

Denote z_{-d} as the set of all z_j for $j \neq d$. Integrating out the mixing proportions P, we can write the conditional distribution of z_d given z_{-d} as follows:

$$p(z_d = i|z_{-d}) = \frac{n_{d,i} + \alpha/N}{D-1+\alpha}, \quad (3)$$

where i indicates each mixture component ranging from 1 to N and $n_{d,i}$ is the number of times that the value of z_j equals to i for $j \neq d$.

If we take $N \rightarrow \infty$ in (3), it is easy to found that the clustering property of the DMA model is the same as that of the DPM model as shown in (2). In [5], it shows that we can choose a reasonable N based on the L_1 distance between the Bayesian marginal density of the data under the DMA model and the DPM model.

```
1 Choose $\gamma_j|\omega \sim \mathrm{B}(1,\omega), j = 1, 2, ..., W$,
2 Choose $|x_d| \sim \mathrm{Poisson}(\xi), d = 1, 2, ..., D$,
3 Choose $G|\lambda \sim \mathrm{DP}(\alpha, G_0)$,
4 Choose $\eta_d|G \sim G, d = 1, 2, ..., D$,
5 Choose $\eta_0|\beta \sim \mathrm{Dirichlet}(\beta_1, \beta_2, ..., \beta_W)$,
6 For $d = 1, 2, ..., D$,
      Choose $x_d\gamma|\eta_d, \gamma \sim$Multinomial$(|x_d|_\gamma; \eta_d)$,
      Choose $x_d(1-\gamma)|\eta_0, \gamma \sim$Multinomial$(|x_d|_{1-\gamma}; \eta_0)$.
```

Fig. 3. The generative process for the DPMFP model.

3.3 Mean Field Variational Inference

Mean field variational inference is a particular class of variational methods [22]. Consider a model with a hyperparameter θ, latent variables $\bar{W} = \{\nu_1, \nu_2, \dots, \nu_S\}$, and data points $\chi = \{x_1, x_2, \dots, x_D\}$. In many situations, the posterior distribution $p(\bar{W}|\chi, \theta)$ is not available in a closed form. The mean field method approximates the posterior distribution $p(\bar{W}|\chi, \theta)$ with a simplified distribution. It starts from a family of distributions Q by using which both the mean field procedure and the subsequent inference procedures are easy to handle. The mean field approximation q is then learned by minimizing the Kullback-Leibler (KL) divergence between the distribution in Q and $p(\bar{W}|\chi, \theta)$ as follow:

$$q = \arg\min_{q^* \in Q} D(q^*(\bar{W}) \| p(\bar{W}|\chi, \theta)), \tag{4}$$

where

$$D(q^*(\bar{W}) \| p(\bar{W}|\chi, \theta)) \\ = E_{q^*}[\log q^*(\bar{W})] - E_{q^*}[\log p(\bar{W}, \chi|\theta)] + \log p(\chi|\theta).$$

Note that since $\log p(\chi|\theta)$ does not depend on the distribution $q^*(\bar{W})$, the minimization of the KL divergence can be cast alternatively as the maximization of a lower bound on the log marginal likelihood as follows:

$$\log p(\chi|\theta) \geq E_{q^*}[\log p(\bar{W}, \chi|\theta)] - E_{q^*}[\log q^*(\bar{W})]. \tag{5}$$

In order to yield a computationally effective inference method, it's very necessary and important to choose a reasonable family of distributions Q. A common and practical method to construct such a family often breaks some of the dependencies between the latent variables. In this paper, we use the fully factorized variational distributions which break all of the dependencies between latent variables.

4 DPMFP and DMAFP Approximation

Formally, we define the following terms:

- A word w is an item from a vocabulary indexed by $\{1, 2, \dots, W\}$.
- A cluster is characterized by a multinomial distribution over words. It is represented by a multinomial parameter.
- A document x is represented as a W-dimensional vector $x_d = \{x_{d1}, x_{d2}, \dots, x_{dW}\}$ where x_{dj} is the number of appearance of the word w_j of the document x_d.
- A document data set χ is a collection of D documents denoted by $\chi = \{x_1, x_2, \dots, x_D\}$.

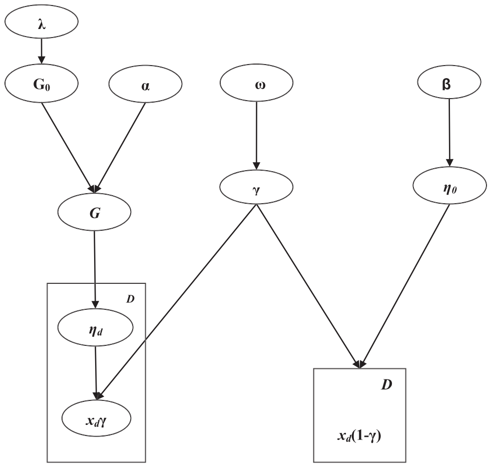

Fig. 4. Graphical representation of the DPMFP model.

We introduce a latent binary vector $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_W\}$ to partition document features into two groups, in particular, the discriminative words and nondiscriminative words. Let Ω denote the discriminative word set. Words not belong to Ω are regarded as nondiscriminative words. For each $j = 1, 2, \dots, W$, we denote

$$\gamma_j = \begin{cases} 1, & \text{if } w_j \in \Omega \\ 0, & \text{if } w_j \notin \Omega. \end{cases} \tag{6}$$

We assign a prior to γ and assume that its elements are independent Bernoulli random variables with common probability distribution $\mathrm{B}(1, \omega)$. The parameter ω can be regarded as the prior probability of each word in the vocabulary which is expected to be discriminative.

Our model assumes the generative process for the document data set χ is as shown in Fig. 3. G_0 is a Dirichlet distribution with parameter $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_W)$; $|x_d|$ is the total appearance of the words in the document x_d; the multinomial parameter η_d represents the specific cluster to which the document x_d belongs; the multinomial parameter η_0 represents the general background sharing by all the documents in the document data set χ; $x_d\gamma$ and $x_d(1-\gamma)$ represent $(x_{d1}\gamma_1, \dots, x_{dW}\gamma_W)$ and $(x_{d1}(1-\gamma_1), \dots, x_{dW}(1-\gamma_W))$, respectively; $|x_d|_\gamma$, which equals to $\sum_{\gamma_j=1} x_{dj}$, is the number of discriminative words in the document x_d; $|x_d|_{1-\gamma}$, which equals to $\sum_{\gamma_j=0} x_{dj}$, is the number of the nondiscriminative words in x_d. In our model, the DP prior is only used for the specific cluster η_d. Note that $|x_d|$ is an ancillary variable as it is independent of all the other data generating parameters. Therefore, we ignore its randomness in the following development. The graphical representation of the DPMFP model is shown in Fig. 4.

We assume that there is no correlation between the set of discriminative words and the set of nondiscriminative words. The conditional probability density function for x_d is given as follows:

$$f(x_d|\gamma, \eta_d, \eta_0) = \frac{|x_d|_\gamma! |x_d|_{1-\gamma}!}{\prod_{j=1}^{W} x_{dj}!} \prod_{j=1}^{W} \eta_{d,j}^{x_{dj}\gamma_j} \eta_{0,j}^{x_{dj}(1-\gamma_j)}. \tag{7}$$

```
Choose $\gamma_j|\omega \sim \mathrm{B}(1,\omega), j = 1,2,...,W$,
Choose $|x_d| \sim \mathrm{Poisson}(\xi), d = 1,2,...,D$,
Choose $\eta_i|\lambda \sim \mathrm{Dirichlet}(\lambda_1,...,\lambda_W), i = 1,2,...,N$,
Choose $\eta_0|\beta \sim \mathrm{Dirichlet}(\beta_1,\beta_2,...,\beta_W)$,
Choose $P|\alpha \sim \mathrm{Dirichlet}(\alpha/N,\alpha/N,...,\alpha/N)$,
Choose $z_d|P \sim \mathrm{Discrete}(p_1,...,p_N), d = 1,2,...,D$,
For $d = 1,2,...,D$,
    Choose $x_d\gamma|\eta_{z_d},\gamma \sim \mathrm{Multinomial}(|x_d|_\gamma;\eta_{z_d})$
    Choose $x_d(1-\gamma)|\eta_0,\gamma \sim \mathrm{Multinomial}(|x_d|_{1-\gamma};\eta_0)$
```

Fig. 5. The generative process for the DMAFP model.

However, since the vocabulary size W is always very large, the law of large numbers and the fact that $\mathrm{E}(\sum_{j=1}^{W}(\eta_{d,j}\gamma_j + \eta_{0,j}(1-\gamma_j)) = 1$ indicate that

$$\sum_{j=1}^{W}(\eta_{d,j}\gamma_j + \eta_{0,j}(1-\gamma_j)) \approx 1. \tag{8}$$

Therefore, we can approximately consider the conditional probability distribution of x_d as a Multinomial distribution with parameters $\{\eta_{d,j}\gamma_j + \eta_{0,j}(1-\gamma_j), j = 1,2,\ldots,W\}$. The approximated probability density function is as follows:

$$f(x_d|\gamma,\eta_d,\eta_0) \approx \frac{|x_d|!}{\prod_{j=1}^{W} x_{dj}!}\prod_{j=1}^{W}\left[\eta_{d,j}^{x_{dj}\gamma_j}\eta_{0,j}^{x_{dj}(1-\gamma_j)}\right]. \tag{9}$$

Furthermore, since the DMA model is a good approximation to the DPM model, we can also adjust our model by applying a DMA prior for the specific cluster of the document. Then, the data set χ can be generated as shown in Fig. 5. N is the number of clusters, the N-dimensional vector P is the mixing proportions for the clusters, and z_d indicates the latent cluster allocation of the document x_d. The graphical representation of the DMAFP model is shown in Fig. 6.

Similar to (9), the probability density function of x_d under the DMAFP model can be approximated as follows:

$$f(x_d|\eta_0,\eta_{z_d},\gamma) \approx \frac{|x_d|!}{\prod_{j=1}^{W} x_{dj}!}\prod_{j=1}^{W}\left[\eta_{z_d,j}^{x_{dj}\gamma_j}\eta_{0,j}^{x_{dj}(1-\gamma_j)}\right]. \tag{10}$$

The introduction of the DMAFP model and the approximation (10) facilitates us to develop effective and fast variational inference algorithm as well as the Gibbs sampling algorithm similar to those for the finite mixture model.

In fact, since Dirichlet distribution is the conjugate prior for the parameter of multinomial distribution, integrating out $\eta_0,\eta_1,\ldots,\eta_N$ in (10), the approximation of the conditional probability density function of the data set χ given $\{z_1,z_2,\ldots,z_D\}$ and γ can be represented as follows:

$$f(\chi|z_1,z_2,\ldots,z_D,\gamma) \approx \prod_{d=1}^{D}\frac{|x_d|!}{\prod_{j=1}^{W} x_{dj}!}\cdot S_{\lambda,\beta}\cdot S_\lambda \cdot S_\beta, \tag{11}$$

where

$$S_{\lambda,\beta} = \left(\frac{\Gamma\left(\sum_{j=1}^{W}\lambda_j\right)}{\prod_{j=1}^{W}\Gamma(\lambda_j)}\right)^N \cdot \frac{\Gamma\left(\sum_{j=1}^{W}\beta_j\right)}{\prod_{j=1}^{W}\Gamma(\beta_j)} \tag{12}$$

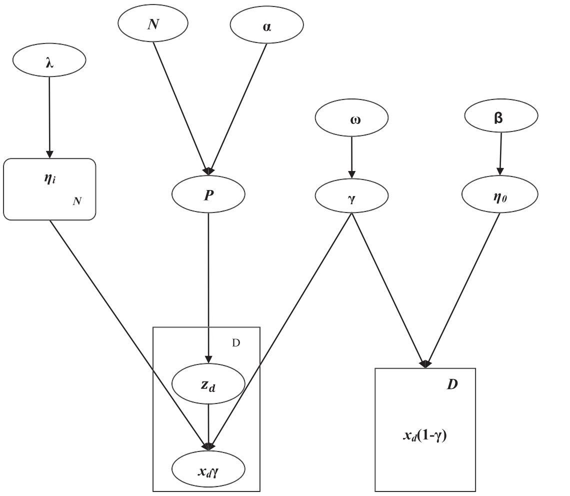

Fig. 6. Graphical representation of the DMAFP model.

$$S_\lambda = \prod_{i=1}^{N}\left(\frac{\prod_{j=1}^{W}\Gamma\left(\lambda_j + \sum_{\{d:z_d=i\}}x_{dj}\gamma_j\right)}{\Gamma\left(\sum_{j=1}^{W}\lambda_j + \sum_{j=1}^{W}\sum_{\{d:z_d=i\}}x_{dj}\gamma_j\right)}\right) \tag{13}$$

$$S_\beta = \frac{\prod_{j=1}^{W}\Gamma\left(\beta_j + \sum_{d=1}^{D}x_{dj}(1-\gamma_j)\right)}{\Gamma\left(\sum_{j=1}^{W}\beta_j + \sum_{j=1}^{W}\sum_{d=1}^{D}x_{dj}(1-\gamma_j)\right)}. \tag{14}$$

5 Algorithm

In this section, we present a variational inference algorithm and a Gibbs sampling algorithm to infer both the clustering structure and the partition of document words simultaneously. Our proposed two algorithms are investigated based on the DMAFP model.

5.1 Variational Inference Algorithm

We use the mean field variational inference algorithm to approximate posterior distribution of the latent variables $\bar{W} = \{\gamma, P, \eta_0, \eta_1, ..., \eta_N, z_1, z_2, ..., z_D\}$ in the DMAFP model. In this setting, the hyperparameter is $\theta = \{\alpha,\omega,\lambda,\beta\}$. It is very natural to choose the mean field variational approximations Q as the following family of distributions:

$$q_\nu(\bar{W}) = q_\sigma(P)q_{\tau_0}(\eta_0)\prod_{j=1}^{W}q_{\omega_j}(\gamma_j)\prod_{i=1}^{N}q_{\tau_i}(\eta_i)\prod_{d=1}^{D}q_{\phi_d}(z_d), \tag{15}$$

where $q_\sigma(P)$ is a Dirichlet distribution with parameter $(\sigma_1,\ldots,\sigma_N)$. $q_{\tau_i}(\eta_i)$ is a Dirichlet distribution with parameter $(\tau_{i1},\ldots,\tau_{iW}), i = 1,2,\ldots,N$. $q_{\omega_j}(\gamma_j)$ is a Bernoulli distribution with parameter $\omega_j, j = 1,2,\ldots,W$. $q_{\phi_d}(z_d)$ is a multinomial distribution with parameter $(\phi_{d,1},\ldots,\phi_{d,N}), d = 1,2,\ldots,D$. In this case, the free variational parameters are

$$\nu = \{\sigma,\omega_1,\ldots,\omega_W,\tau_0,\tau_1,\ldots,\tau_N,\phi_1,\ldots,\phi_D\}. \tag{16}$$

In order to acquire a good approximation for the posterior distribution of the latent variables $\bar{W}$, we need to iteratively update the free variational parameters ν and maximize the lower bound on the log marginal likelihood as follows:

$$\log p(\chi|\theta) \geq E_{q_\nu}[\log f(\bar{W},\chi|\theta)] - E_{q_\nu}[\log q_\nu(\bar{W})]. \tag{17}$$

Therefore, the lower bound of the log marginal likelihood is as follows:

$$\begin{aligned} L &= E_{q_\nu}[\log f(\bar{W}, \chi|\theta)] - E_{q_\nu}[\log q_\nu(\bar{W})] \\ &= E_{q_\nu}[\log f(\chi|\bar{W}, \theta)] + E_{q_\nu}[\log f(\bar{W}|\theta)] \\ &\quad - E_{q_\nu}[\log q_\nu(\bar{W})]. \end{aligned} \tag{18}$$

To maximize the lower bound L, the update equations for ν are as follows:

$$\sigma_i = \frac{\alpha}{N} + \sum_{d=1}^{D} \phi_{d,i} \tag{19}$$

$$\tau_{0j} = \beta_j + \sum_{d=1}^{D} x_{dj}(1-\omega_j) \tag{20}$$

$$\tau_{ij} = \lambda_j + \sum_{d=1}^{D} x_{dj}\omega_j\phi_{d,i} \tag{21}$$

$$\phi_{d,i} = \exp\{M_{d,i}\} \tag{22}$$

$$\omega_j = \frac{\exp(A_j)}{1+\exp(A_j)}, \tag{23}$$

where $i \in \{1, \ldots, N\}, j \in \{1, \ldots, W\}, d \in \{1, \ldots, D\}$ and

$$\begin{aligned} M_{d,i} &= \sum_{j=1}^{W} x_{dj}\omega_j \left(\psi(\tau_{ij}) - \psi\left(\sum_{j=1}^{W} \tau_{ij} \right) \right) \\ &\quad + \psi(\sigma_i) - \psi\left(\sum_{i=1}^{N} \sigma_i \right) - 1 \end{aligned} \tag{24}$$

$$\begin{aligned} A_j &= \log \frac{\omega}{1-\omega} - \sum_{d=1}^{D} x_{dj} \left(\psi(\tau_{0j}) - \psi\left(\sum_{j=1}^{W} \tau_{0j} \right) \right) \\ &\quad + \sum_{d=1}^{D}\sum_{i=1}^{N} x_{dj}\phi_{d,i} \left(\psi(\tau_{ij}) - \psi\left(\sum_{j=1}^{W} \tau_{ij} \right) \right). \end{aligned} \tag{25}$$

The digamma function is denoted by ψ which arises from the derivative of the log normalization factor in the Dirichlet distribution.

Repeatedly updating the variational parameter ν through (19) to (23) would increase the low bound L and finally acquire a local maxima of L. The proof is shown in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.27.When the improvement of L is less than a threshold, we estimate the latent clustering structure and the partition of document words by the variational parameter $\{\phi_{d,i}, d = 1, \ldots, D, i = 1, \ldots, N\}$ and $\{\omega_1, \ldots, \omega_W\}$, respectively. The cluster to which the document x_d belongs is determined by the value of $\phi_{d,i}$. In particular, let $\phi_{d,i}$ be the largest value acquired by the document x_d, x_d will then be assigned to the cluster labeled by i. The word w_j is discriminative when ω_j is larger than a threshold ε. (We set ε as 0.7 in our experiments.) Otherwise, w_j is regarded as nondiscriminative.

Note that as the traditional EM algorithm, our proposed mean field variational inference method yields a local maxima. For practical applications, we run the algorithm multiple times with different initial values. Some authors suggested that the inference should based on the result which acquires the largest L as shown in (18) [22]. However, for our proposed variational inference algorithm designed for the document clustering, our large amounts of experiments indicate that an effective way to choose a good result is to choose the one which acquires the largest value of $E_{q_\nu}[\log f(\bar{W}, \chi|\theta)]$. Since $E_{q_\nu}[\log f(\bar{W}, \chi|\theta)] = E_{q_\nu}[\log f(\bar{W}|\chi, \theta)] + \log f(\chi|\theta)$, this method tends to choose the result which reaches the highest posterior likelihood of the latent variables $\bar{W}$. The calculation of $E_{q_\nu}[\log f(\bar{W}, \chi|\theta)]$ is shown in the Appendix, available in the online supplemental material.

5.2 Blocked Gibbs Sampling Algorithm

Another effective inference algorithm for our proposed model is the blocked Gibbs sampling algorithm. Many authors have used this method to infer various model based on the DP prior. For the DMAFP model, the state of the Markov chain is $\bar{W} = \{\gamma, P, \eta_0, \eta_1, \ldots, \eta_N, z_1, \ldots, z_D\}$. Let $\{z_1^*, \ldots, z_M^*\}$ denote the set of distinct values of $\{z_1, z_2, \ldots, z_D\}$. After initializing the latent variable $\{\gamma_1, \gamma_2, \ldots, \gamma_W, z_1, \ldots, z_D\}$ and the hyperparameter θ, the blocked Gibbs sampling procedure iterates between the following steps:

1. Update the latent discriminative words indicator γ by repeating the following Metropolis step R times: A new candidate γ_{new} which adds or deletes a discriminative word is generated by randomly picking one of the W indices in γ_{old} and changing its value. The new candidate is accepted with the probability

$$\min\left\{1, \frac{f(\gamma_{new}|\chi, z)}{f(\gamma_{old}|\chi, z)}\right\}, \tag{26}$$

   where $f(\gamma|\chi, z) \propto f(\chi|z, \gamma)p(\gamma)$ and $f(\chi|z, \gamma)$ is given by (11).

2. Conditioned on other latent variables, for $i = 1, 2, \ldots, N$, if i is not in $\{z_1^*, z_2^*, \ldots, z_M^*\}$, draw η_i from a Dirichlet distribution with parameter λ. Otherwise, update η_i by sampling a value from a Dirichlet distribution with parameter:

$$\left\{ \lambda_1 + \sum_{\{d:z_d=i\}} x_{d1}\gamma_1, \ldots, \lambda_W + \sum_{\{d:z_d=i\}} x_{dW}\gamma_W \right\}. \tag{27}$$

3. Update η_0 by sampling a value from a Dirichlet distribution with parameter:

$$\left\{ \beta_1 + \sum_{d=1}^{D} x_{d1}(1-\gamma_1), \ldots, \beta_W + \sum_{d=1}^{D} x_{dW}(1-\gamma_W) \right\}. \tag{28}$$

4. Update P by sampling a value from a Dirichlet distribution with parameter:

$$\left\{ \frac{\alpha}{N} + \sum_{d=1}^{D} I(z_d = 1), \ldots, \frac{\alpha}{N} + \sum_{d=1}^{D} I(z_d = N) \right\}, \tag{29}$$

where $I(z_d = i)$ is an indicator function which equals to 1 if $z_d = i$.

5. Conditioned on other latent variables, for $d = 1, 2, \dots, D$, update z_d by sampling a value from a discrete distribution with parameter $\{s_{d1}, s_{d2}, \dots, s_{dN}\}$, where

$$\sum_{i=1}^{N} s_{di} = 1 \quad \text{and} \quad s_{di} \propto p_i f(x_d|\eta_i, \eta_0, \gamma), i = 1, \dots, N. \tag{30}$$

After the Markov chain has reached its stationary distribution, we collect H samples of $\{z_1, \dots, z_D\}$ and $\{\gamma_1, \dots, \gamma_W\}$. Latent document labels and the partition of document words are then estimated as follows:

1. The estimated label of document x_d is the most frequent value of z_d in the last H samples;
2. The word w_j is discriminative if the average value of the last H sample of γ_j is bigger than a threshold ε. (We set ε as 0.7 in our experiments.) Otherwise, w_j is regarded as nondiscriminative.

5.3 Update for the Hyperparameter θ

In the above two algorithms, the hyperparameter $\theta = \{\alpha, \omega, \lambda, \beta\}$ is assumed to be unchanged. However, in some situations, the parameter θ affects the inference result. Different setting of θ may lead to different inference results. Therefore, it is useful to find appropriate value to θ.

In our approach, we investigate an update method to learn the value of θ. First, θ is initialized by an arbitrary value. The value of θ will be updated to an appropriate value in the process of document clustering by maximizing the lower bound of L with respect to θ. In the variational Bayes, this is often called variational M-step and (19)-(23) are called the variational E-step. Setting the partial derivative with respect to θ to be zero, we obtain the following equations for updating $\hat{\theta}$:

$$\omega = \frac{\sum_{j=1}^{W} \omega_j}{W} \tag{31}$$

$$\psi(\hat{\alpha}) - \psi\left(\frac{\hat{\alpha}}{N}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\psi\left(\sum_{n=1}^{N}\sigma_n\right) - \psi(\sigma_i)\right) \tag{32}$$

$$\psi\left(\sum_{j=1}^{W}\hat{\lambda}_j\right) - \psi(\hat{\lambda_s}) = \frac{1}{N}\sum_{i=1}^{N}\left(\psi\left(\sum_{j=1}^{W}\tau_{ij}\right) - \psi(\tau_{is})\right) \tag{33}$$

$$\psi\left(\sum_{j=1}^{W}\hat{\beta}_j\right) - \psi(\hat{\beta}_s) = \psi\left(\sum_{j=1}^{W}\tau_{0j}\right) - \psi(\tau_{0s}), \tag{34}$$

where $s = 1, 2, \dots, W$.

For the blocked Gibbs sampling method, one common method for updating the hyperparameter θ is to assign a prior $p(\theta)$ to θ and resample θ from its posterior distribution. More discussion about this update method can be found in [5] and [6].

6 Experiment

We study the performance of our proposed approach by two sets of experiments. For the first set of experiments, a synthetic data set is used. For the second set of experiments, our proposed approach is evaluated via real document data sets.

6.1 Evaluation Metric

The normalized mutual information (NMI) is used to evaluate the quality of a clustering solution. NMI is an external clustering validation metric that effectively measures the amount of statistical information shared by the random variables representing the cluster assignments and the user-labeled class assignments of the data points. In general, NMI is estimated as follows [23]:

$$NMI = \frac{\sum_{h,l} d_{hl} \log\left(\frac{D \cdot d_{hl}}{d_h c_l}\right)}{\sqrt{\left(\sum_h d_h \log\left(\frac{d_h}{D}\right)\right)\left(\sum_l c_l \log\left(\frac{c_l}{D}\right)\right)}}, \tag{35}$$

where D is the number of documents, d_h is the number of documents in class h, c_l is the number of documents in cluster l, and d_{hl} is the number of documents in class h as well as in cluster l. The NMI value is 1 when a clustering solution perfectly matches the user-labeled class assignments and close to 0 for a random document partitioning.

6.2 Synthetic Data Set Experiments

6.2.1 Experimental Data Set

The synthetic data set consists of 600 data points with 2,000 features. Data points were generated by two different processes with seven multinomial distributions. Six of them are used in the first process to generate discriminative features. The remaining one is used in the second process to generate nondiscriminative features. In the first process, a multinomial mixture model with six components is used to model six different clusters. Each component of the multinomial mixture model represents one cluster parameterized by one multinomial distribution parameter. Each cluster contains 100 data points. For each data points, the first 200 features were regarded as discriminative features generated from one of the six components. The second process was used to generate nondiscriminative features. In particular, the remaining 1,800 features were regarded as nondiscriminative features generated from one multinomial distribution. The data were generated as follows:

$$\begin{aligned}(x_{i1}, \dots, x_{i200}) &\sim \text{Multinomial}(\pi_j; 200), j = 1, \dots, 6,\\ &\quad i = 1 + 100(j-1), \dots, 100j, (x_{i201}, \dots, x_{i2000})\\ &\quad \sim \text{Multinomial}\big(\pi^*; 200\big), i = 1, \dots, 600,\end{aligned}$$

where $(\pi^*, 200)$ and $(\pi_j, 200), j = 1, \dots, 6$ are different multinomial parameters. In our experiment, $\pi^*, \pi_1, \dots, \pi_6$ are chosen randomly.

6.2.2 Experimental Setup

Both the variational inference algorithm and the blocked Gibbs sampling algorithm were used to infer the cluster structure as well as the partition of features for the synthetic data set generated from the above process. In the variational

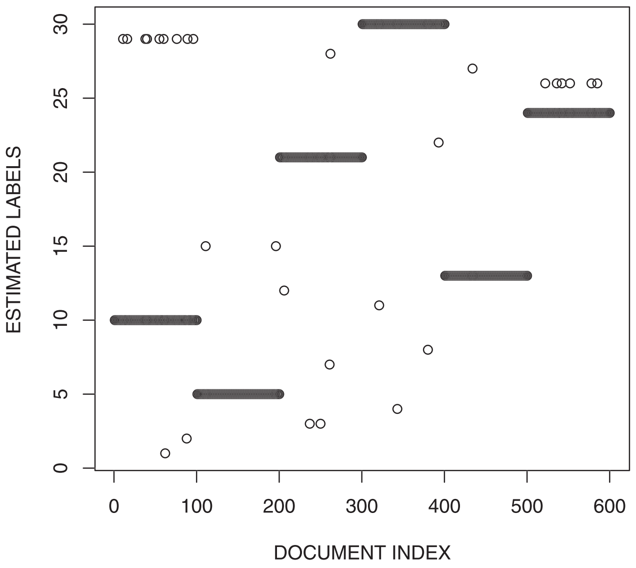

Fig. 7. Estimated labels of the synthetic data set acquired by the variational inference algorithm.

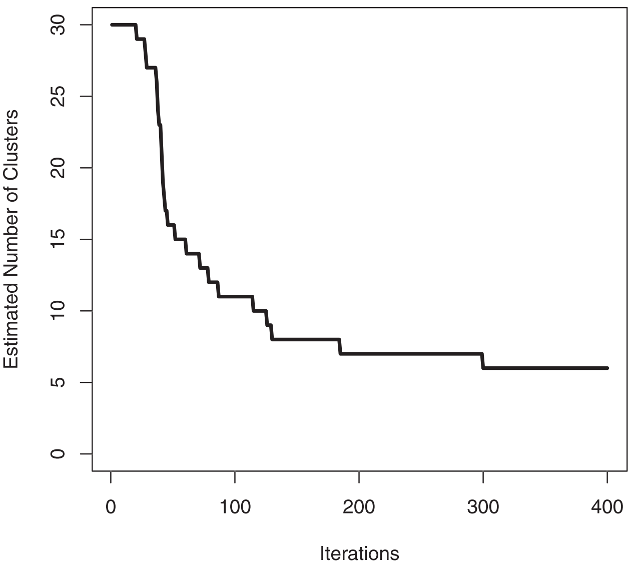

Fig. 8. Trace plot for the number of clusters for the synthetic data set acquired by the Gibbs sampling algorithm.

inference algorithm, we set $N = 30, \alpha = 1.0$, $\omega = 0.01$, $\lambda_j = 0.02$, and $\beta_j = 4.0$, where $j = 1, 2, \ldots, 2{,}000$. Only the variational parameter $\{\phi_{d,i}, d = 1, \ldots, 600, i = 1, \ldots, 50\}$ and $\{\omega_j, j = 1, \ldots, 2{,}000\}$ need to be initialized. We initialized the parameter ϕ randomly while the initial values for $\{\omega_j, j = 1, \ldots, 2{,}000\}$ are set to 1. The algorithm was run 20 times. Each time the process was stopped when the relative change in the log marginal probability bound (17) is less than $1e^{-7}$.

For the blocked Gibbs sampling algorithm, we used the same setting for the parameter N and $\{\alpha, \omega, \lambda, \beta\}$ as the variational inference algorithm. The Metropolis step R was chosen to be 200. The initial labels $\{z_d, d = 1, \ldots, 600\}$ were chosen randomly from $\{1, 2, \ldots, 50\}$ and only one feature was chosen randomly to be discriminative. We also ran the algorithm 20 times. Each time, we ran 400 samplings in which only the last 100 samples were used to infer latent data labels as well as the partition of features.

6.2.3 Experimental Results for the DMAFP Model

The NMI acquired by the variational inference algorithm is 0.945. Fig. 7 depicts estimated labels of data points of the synthetic data set. As shown in Fig. 7, data points are apparently partitioned into six clusters which provide strong support for the six true clusters indicated by the discriminative features. There are a few small clusters discovered. The reason is that the synthetic data set is generated arbitrarily. Outlier data points are unavoidably generated which are regarded as more likely belonging to new clusters. In respect to the inference of the discriminative features set, the result of the variational inference algorithm indicates that there are 375 discriminative features including the true 200 discriminative features.

The averaged NMI acquired by the blocked Gibbs sampling algorithm is 0.979. The partition of features estimated via this algorithm is exactly the same as the true one. The trace plots for the number of clusters and the number of discriminative features are shown in Figs. 8 and 9, respectively. The sampling algorithm was stable after about 300 iterations.

The above experimental results indicate that our proposed two algorithms are effective for the task of clustering when there are a large number of nondiscriminative features. Compared with the variational inference algorithm, the blocked Gibbs sampling algorithm could acquire slightly higher accuracy. However, the blocked Gibbs sampling algorithm is time consuming. It took almost 3 hours for the blocked Gibbs sampling algorithm to finish 20 runs while the variational inference algorithm only took 10 minutes. Furthermore, it is very difficult for user to diagnose the convergence of the Gibbs sampling method. As shown in Fig. 8, the decrement of the number of clusters turns to be very slow when the estimated number of clusters approaches the true value. The variational inference algorithm is fast, deterministic, and acquire comparable cluster-

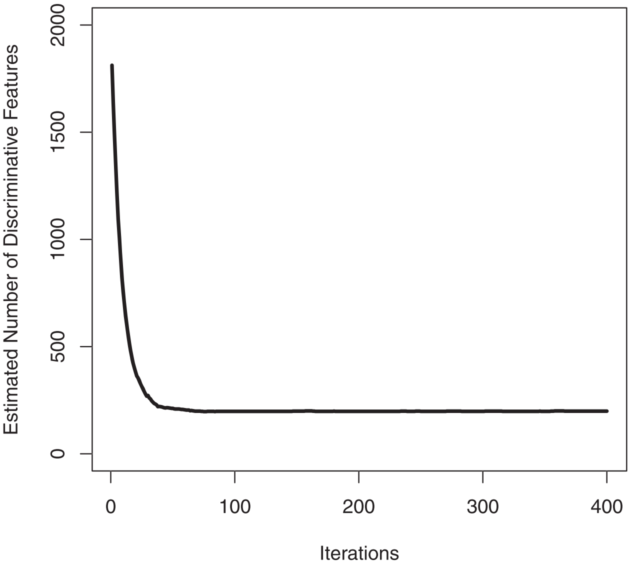

Fig. 9. Trace plot for the number of discriminative features for the synthetic data set acquired by the Gibbs sampling algorithm.

TABLE 1
Real Document Data Sets Description (D: Number of Documents, K: Number of Clusters, W: Vocabulary Size)

Dataset	D	K	W
News-different-3	300	3	2121
News-similar-3	300	3	1767
Classic 400	400	3	6205
News-all-20	17,820	20	50,799

ing solution. We use the variational inference algorithm to analyze real document data sets in the following section.

6.3 Real Data Set Experiments

6.3.1 Experimental Data Sets

Four real document data sets were used for evaluating our proposed approach, in particular, *News-different-3*, *News-similar-3*, *News-all-20*, and *Classic 400*. The summary of these four text document data sets is shown in Table 1. The first three data sets were derived from the *20-Newsgroups* collection.[1] This collection contains messages collected from 20 different Usenet newsgroups, about 1,000 messages from each newsgroup. *News-different-3* consists of 300 messages from three newsgroups on relatively different topics (alt.atheism, rec.sport.baseball, sci.space) with well-separated clusters. *News-similar-3* consists of 300 messages from three newsgroups on similar topics (comp.graphics, comp.os.ms-windows, comp.windows.x) where cross-posting often occurs. The *News-all-20* data set is a large data set which contains nearly all messages in the *20-Newsgroups* collection from 20 different newsgroups. We only filtered out substantially short messages with the number of words smaller than 30 because these messages are too short and usually have no overlap with other messages. The *Classic 400*, which is a typical unbalanced data set, is used by the EDCM model proposed in [1].

We preprocessed the *News-different-3*, *News-similar-3*, *News-all-20*, and *Classic 400* data sets by stopword removal. High-frequency and Low-frequency words were removed following the methodology presented in [24]. The purpose of such processing is to eliminate those words which obviously do not define the latent cluster structure. Thresholds for removing high-frequency and low-frequency words for *News-different-3* and *News-similar-3* data sets were set 100 and 1, respectively. Thresholds for removing high-frequency and low-frequency words for *Classic400* data set were set to 200 and 1, respectively. For the *News-all-40* data set, thresholds were set to 4,000 and 1.

6.3.2 Experimental Setup

We evaluate our proposed approach, namely DMAFP, based on the variational inference algorithm. We set $\alpha = 5.0, \omega = 0.01, \lambda = 3.0, \beta = 0.5$ for all experiments. The setting of initial values of these hyperparameters is arbitrary because all of them are updated during the clustering process by the method proposed in Section 5.3. We set N to 30 for the data sets *News-different-3*, *News-similar-3*, and *Classic 400*. N was set to 80 for the *News-all-20* data set in consideration of its large number of documents. For each data set, we conducted experiments 20 times and chose the result which acquired the largest value of $E_{q_\nu}[\log f(\bar{W}, \chi|\theta)]$ as discussed in Section 5.1.

TABLE 2
Comparison of the Document Clustering Performance on the *News-Different-3*, the *News-Similar-3*, and the *Classic 400* Data Sets

Approach	News-different-3	News-similar-3	Classic400
DMAFP	0.801	0.261	0.747
DMA	0.724	0.093	0.690
EM-MN(K=2)	0.464	0.076	0.361
EM-MN(K=3)	0.867	0.081	0.496
EM-MN(K=10)	0.634	0.126	0.506
K-MEANS(K=2)	0.256	0.043	0.354
K-MEANS(K=3)	0.435	0.106	0.397
K-MEANS(K=10)	0.180	0.067	0.247
LDA(K=2)	0.557	0.104	0.503
LDA(K=3)	0.617	0.131	0.558
LDA(K=10)	0.532	0.114	0.420
EDCM(K=2)	0.386	0.085	0.243
EDCM(K=3)	0.734	0.163	0.684
EDCM(K=10)	0.561	0.228	0.403

$\mathrm{EDCM}(\mathrm{K}=2,3,10)$ *(or* $\mathrm{EM\text{-}MN}(\mathrm{K}=2,3,10)$, $\mathrm{K\text{-}MEANS}(\mathrm{K}=2,3,10)$, $\mathrm{LDA}(\mathrm{K}=2,3,10)$*) indicates the* EDCM *(or* EM-MN, K-MEANS, LDA*) algorithm with the predefined number of clusters* $K = 2; 3; 10$.

For comparative investigation, a standard model-based clustering approach [3], labeled as EM-MN, and the K-MEANS approach, labeled as K-MEANS, were investigated as benchmark. We also ran experiments for the LDA approach [25], labeled as LDA, and a state-of-the-art model-based clustering approach [1], labeled as EDCM. All of these approaches assume that the number of clusters is predefined. For the LDA approach, the initial value of α_{LDA} was set to 1 and was then estimated along the document clustering process. β_{LDA} was initialized randomly. Each document is associated with the cluster to which the value of γ_{LDA} is maximum. Since the deterministic annealing procedure [10] allows EM to find better local optima of the likelihood function and therefore improve the clustering quality, we investigated it to the EM-MN and EDCM approaches. We also studied the performance of these approaches when incorrect numbers of clusters are provided. To evaluate the effectiveness of partitioning document discriminative words and nondiscriminative words for document clustering, we also investigate our proposed approach without feature partition, labeled as DMA. The variational inference algorithm was used to infer the document collection structure. Each approach was conducted 20 times. The performance was computed by taking the average of these 20 experiments.

6.3.3 Experimental Results for the DMAFP Model on the News-Different-3, *the* News-Similar-3, *and the* Classic 400 *Data Sets*

Table 2 depicts the document clustering performance acquired by the DMAFP, DMA, EM-MN, K-MEANS, LDA, and EDCM approaches on the *News-different-3*, the *News-simlar-3*, and the *Classic 400* data sets. The experimental results show that our proposed DMAFP approach

[1] The description of the *20-Newsgroups* data set can be found at http://people.csail.mit.edu/jrennie/20Newsgroups.

TABLE 3
Estimated Number of Clusters on the *News-Different-3*, the *News-Similar-3*, and the *Classic 400* Data Sets

Approach	News-different-3	News-similar-3	Classic400
DMAFP	5	6	8
DMA	9	9	17

achieves better performance compared with the state-of-the-art approaches. The DMAFP model is useful for discovering document collection structure. When correct number of clusters K is provided, the clustering performances achieved by the LDA, the K-MEANS, the EDCM and the EM-MN approaches are comparable to our approach ($K = 3$ for these three data sets). When the number of clusters K was given imprecise, the performances of the EM-MN, the K-MEANS, the LDA, and the EDCM approaches are much worse in most of situations. Therefore, having the correct number of K is crucial to these approaches. On the contrary, our proposed DMAFP model is robust for document clustering without the necessary to have the number of clusters known in advance. Furthermore, the DMA approach also shows promising results in the experiments on the *News-different-3* and the *Classic 400* data sets. It performs much worse for the *News-similar-3* data set. The reason is that documents in the *News-different-3* and the *Classic 400* data sets are relatively well separated and there are many discriminative words aiding the clustering process. However, the *News-similar-3* data set contains similar clusters with a large number of nondiscriminative words which results in unclear document collection structure. When the number of clusters is unknown, the document collection structure becomes more unclear. An improper estimation of K easily misleads document clustering which leads to even worse document partition results in return. Therefore, partitioning discriminative words and nondiscriminative words is useful for document clustering especially when the number of clusters is unknown.

Table 3 shows the number of clusters estimated by our proposed DMAFP approach. The DMA approach is also investigated for comparison analysis. From Table 3, it shows that our estimation for the number of clusters are relatively bigger than the true one. The reason is that there are a number of outlier documents in the real document data set. These outlier documents are dissimilar with other documents belonging to the same cluster and are regarded as belonging to new clusters in the DMAFP approach. The same effect could be easily achieved when documents are manually partitioned into groups. The DMAFP approach acquires more precise estimation compared with the DMA approach. Therefore, partitioning discriminative words and nondiscriminative words is useful for estimating the number of clusters K.

6.3.4 Experimental Result on Large Real Document Data Sets

We tested our proposed DMAFP model on the *News-all-20* data set for evaluating our proposed approach on large real document data sets. Experiments on the K-MEANS, the EM-MN, the LDA, and the EDCM approaches were also conducted with correct and incorrect number of clusters.

TABLE 4
Comparison of the Document Clustering Performance on the *News-All-20* Data Set and the *News-Complete-20* Data Set

Approach	News-all-20	News-complete-20
DMAFP	0.534	0.510
DMA	0.512	0.497
EM-MN(K=10)	0.486	0.439
EM-MN(K=20)	0.531	0.513
EM-MN(K=50)	0.507	0.489
K-MEANS(K=10)	0.184	0.181
K-MEANS(K=20)	0.229	0.205
K-MEANS(K=50)	0.195	0.182
LDA(K=10)	0.482	0.463
LDA(K=20)	0.559	0.532
LDA(K=50)	0.508	0.492
EDCM(K=10)	0.401	0.339
EDCM(K=20)	0.510	0.432
EDCM(K=50)	0.489	0.388

$\mathrm{EDCM}(\mathrm{K} = 10, 20, 50)$ *(or* $\mathrm{EM\text{-}MN}(\mathrm{K} = 10, 20, 50)$, $\mathrm{K\text{-}MEANS}(\mathrm{K} = 10, 20, 50)$, $\mathrm{LDA}(\mathrm{K} = 10, 20, 50)$*) indicates the* EDCM *(or* EM-MN, K-MEANS, LDA*) algorithm with the predefined number of clusters* $\mathrm{K} = 10; 20; 50$.

Besides directly evaluating our proposed approach, we also conducted experiments investigating the effect of the document word partition. Recall that we preprocessed the *News-all-20* data set by removing stopwords and high-frequency/low-frequency words. The motivation of this preprocessing process is to remove obviously nondiscriminative words. To investigate the effect of nondiscriminative words, we derived another data set, *News-complete-20*, without this preprocessing process. We removed those words which only appear in one document as they are obviously not useful for document clustering. The *News-complete-20* data set consists of the same set of documents with the *News-all-20* data set but with a relatively larger set of words. Experiments conducted on the *News-all-20* and *News-complete-20* data sets were depicted in Table 4.

From Table 4, the DMAFP approach achieves better performance than most of other approaches. Experimental results on the *News-all-20* data set depict that the DMAFP approach achieves comparable results with the EM-MN, the LDA, and the EDCM approaches and better results compared with the K-MEANS approach. Although the LDA approach achieves slightly better result, it needs the number of clusters as the predefined parameter. When the number of clusters is incorrect, all approaches perform worse compared with the DMAFP approach. Therefore, the DMAFP approach is effective for large real data sets. The DMA approach also shows promising results but performs much worse compared with DMAFP approach which indicates the usefulness of feature partition to document clustering. From experimental results on the *News-complete-20* data set, all approaches achieve slightly worse performance. Therefore, nondiscriminative words confuse the clustering process. Moreover, the differences between experimental results of these two data sets are small. The reason is that the *News-complete-20* data set only includes 234 more words than the *News-all-20* data set. Compared with the whole set of 50,799 words in the *News-all-20* data set, the effect of these 234 words is small to the document clustering.

TABLE 5
Estimated Number of Clusters on the *News-All-20* and the *News-Complete-20* Data Sets

Approach	News-all-20	News-complete-20
DMAFP	52	53
DMA	55	55

Table 5 shows the number of clusters estimated by our proposed DMAFP approach. The DMA approach is also investigated for comparison analysis. All estimations on the number of clusters are relatively bigger than the true one because of the existence of outlier documents. The DMAFP approach acquires more precise estimation compared with the DMA approach. Therefore, partitioning discriminative words and nondiscriminative words is useful for document clustering.

We also investigated the partition of document words discovered by the DMAFP approach. It estimated that there were 43,701 discriminative words and 7,098 nondiscriminative words. Our further analysis on the word distribution found that about 82 percent of the 43,701 discriminative words appeared in at most three topics in the data set. It indicates that most of these words are actually effective for the clustering process. We discovered 3,098 nondiscriminative terms, of which 5,541 were related to at least eight topics. Hence, our approach could effectively partition document words and is effective for the document clustering for large document data sets.

7 Conclusions and Future Work

In this paper, we proposed an approach which handles document clustering and feature partition simultaneously. A document clustering approach is investigated based on the DPM model which groups documents into an arbitrary number clusters. Document words are partitioned according to their usefulness to discriminate the document clusters. The discriminative words are used to determine the document collection structure. Nondiscriminative words are regarded to be generated from a general background shared by all documents. Both the variational inference algorithm and the blocked Gibbs Sampling technique are proposed to infer the cluster structure as well as the latent discriminative word subset. Our experiment shows that our approach acquires high clustering accuracy and reasonable partition of document words. The comparison between our approach and state-of-the-art approaches indicates that our approach is robust and effective for document clustering. Our analysis of the experiment result also shows that the DPM model with automatic feature partition method could effectively discover word partitions and improve the document clustering quality.

For future research, an interesting direction is to study how to adapt our proposed approach for the semi-supervised document clustering. With more and more labeled documents or constraints are available in real life, the additional information could be used to improve the performance of our approach from at least two aspects. On the one hand, the additional information can be used to select good model parameters. On the other hand, it could be used to guide our model select more precise discriminative words set.

Acknowledgments

The work described in this paper is substantially supported by the National Natural Science Foundation of China (Grant Nos: 61202089, 11071128, and 11131002), the Science and Technology Fund of Guizhou Province (Grant No: 2172), the Hong Kong Polytechnic University Internal Competitive Research Grant (Grant No: A-PJ72), and the Doctoral Fund of Ministry of Education of China (Grant No: 20110031110002).

References

[1] C. Elkan, "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution," *Proc. Int'l Conf. Machine Learning,* pp. 289-296, 2006.

[2] R. Madsen, D. Kauchak, and C. Elkan, "Modeling Word Burstiness Using the Dirichlet Distribution," *Proc. Int'l Conf. Machine Learning,* pp. 545-552, 2005.

[3] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchel, "Text Classification from Labeled and Unlabeled Documents Using Em," *J. Machine Learning,* vol. 39, no. 2, pp. 103-134, 2000.

[4] C. Antoniak, "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics,* vol. 2, no. 6, pp. 1152-1174, 1974.

[5] J. Ishwaran and L. James, "Gibbs Sampling Methods for Stick-Breaking Priors," *J. Am. Statistical Assoc.,* vol. 96, no. 453, pp. 161-174, 2001.

[6] R. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *J. Computational and Graphical Statistics,* vol. 9, no. 2, pp. 249-265, 2000.

[7] D. Blackwell and J. MacQueen, "Ferguson Distribution via Polya URN Schemes," *The Annals of Statistics,* vol. 1, no. 2, pp. 353-355, 1973.

[8] T. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics,* vol. 1, no. 2, pp. 209-230, 1973.

[9] M.H.C. Law, M.A.T. Figueiredo, and A.K. Jain, "Simultaneous Feature Selection and Clustering Using Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1154-1166, Sept. 2004.

[10] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems," *Proc. IEEE,* vol. 86, no. 11, pp. 2210-2239, Nov. 1998.

[11] P. Smyth, "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood," *Statistics and Computing,* vol. 10, no. 1, pp. 63-72, 2000.

[12] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freedman, "Autoclass: A Bayesian Classification System," *Proc. Int'l Conf. Machine Learning,* pp. 54-64, 1988.

[13] J. Rissanen, "Modeling by Shortest Data Description," *Automatica,* vol. 14, pp. 465-471, 1978.

[14] C. Wallace and P. Freedman, "And Inference by Compact Coding," *J. Royal Statistical Soc., Series B,* vol. 49, no. 3, pp. 240-265, 1987.

[15] H. Bozdogan, "Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model Using Model-Selection Criteria," Technical Report UIC/DQM/A83-1, Quantitative Methods Dept., Univ. of Illinois, Chicago, IL, 1983.

[16] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics,* vol. 6, no. 2, pp. 461-464, 1978.

[17] C. Fraley and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer J.,* vol. 41, no. 8, pp. 578-588, 1998.

[18] G. Yu, R. Huang, and Z. Wang, "Document Clustering via Dirichlet Process Mixture Model with Feature Selection," *Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining,* pp. 763-772, 2010.

[19] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet Processes," *J. Am. Statistical Assoc.*, vol. 101, no. 476, pp. 1566-1581, 2007.

[20] P.J. Green and S. Richardson, "Modelling Heterogeneity with and without the Dirichlet Process," *Scandinavian J. Statistics,* vol. 28, no. 2, pp. 355-377, 2001.

[21] H. Ishwaran and M. Zarepour, "Exact and Approximate Sum-Representations for the Dirichlet Process," *Canadian J. Statistics,* vol. 30, no. 2, pp. 269-283, 2002.

[22] D. Blei and M. Jordan, "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis,* vol. 1, no. 1, pp. 121-144, 2006.

[23] S. Zhong, "Semi-Supervised Model-Based Document Clustering: A Comparative Study," *J. Machine Learning,* vol. 65, no. 1, pp. 3-29, 2006.

[24] I.S. Dhillon and D.S. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," *J. Machine Learning,* vol. 42, no. 1, pp. 143-175, 2001.

[25] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

Ruizhang Huang received the BS degree in computer science from Nankai University, China, in 2001 and the Mphil and PhD degrees in the systems engineering and engineering management from the Chinese University of Hong Kong, Hong Kong, in 2003 and 2008, respectively. In 2007, she joined the Hong Kong Polytechnic University, Hong Kong, as a lecturer. Since 2011, she has been with Guizhou University as an associate professor. She is an active researcher in the areas of data mining, text mining, machine learning, and information retrieval. She has published a number of papers that have been published in prestigious journals and conferences.

Guan Yu received the BS degree in mathematics and the MS degree in statistics from Nankai University, China. He is currently working toward the PhD degree in the Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill. His research interests include statistical machine learning, nonparametric statistics, and functional estimation.

Zhaojun Wang received the BS degree from the Department of Mathematics, Nankai University, in 1987, the master's degree from the Department of Mathematical Statistics, East China Normal University, in 1990, and the PhD degree at the School of Mathematical Sciences, Nankai University, in 1995. He is currently a statistical professor at Nankai University. His research interests include statistical process control, variable section, dimension reduction, and high-dimensional data analysis. He has published more than 70 refereed papers.

Jun Zhang received the PhD degree in electrical engineering from the City University of Hong Kong, Kowloon, Hong Kong, in 2002. From 2003 to 2004, he was a Brain Korean 21 postdoctoral fellow in the Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. Since 2004, he has been with Sun Yat-Sen University, Guangzhon, China, where he is currently a professor and the PhD supervisor in the Department of Computer Science. He has authored seven research books and book chapters, and more than 100 technical papers in his research areas. His research interests include computational intelligence, data mining, operations research, wireless sensor networks, and power electronic circuits.

Liangxing Shi received the BS and PhD degrees from the School of Management, Tianjin University, China, in 2000 and 2008, respectively. He is currently an associate professor in the School of Management, Tianjin University. His main research interests include industrial engineering, quality control, and data mining.