Document Clustering via Dirichlet Process Mixture Model with Feature Selection

Guan Yu
Dept. of Industrial and Systems Engineering
The Hong Kong Polytechnic University
Kowloon, Hong Kong
yu.guan@polyu.edu.hk

Ruizhang Huang
Dept. of Industrial and Systems Engineering
The Hong Kong Polytechnic University
Kowloon, Hong Kong
mfrzh@polyu.edu.hk

Zhaojun Wang
Dept. of Statistics
School of Mathematical Sciences
Nankai University
Tianjin, China
zjwang@nankai.edu.cn

ABSTRACT

One essential issue of document clustering is to estimate the appropriate number of clusters for a document collection to which documents should be partitioned. In this paper, we propose a novel approach, namely DPMFS, to address this issue. The proposed approach is designed 1) to group documents into a set of clusters while the number of document clusters is determined by the Dirichlet process mixture model automatically; 2) to identify the discriminative words and separate them from irrelevant noise words via stochastic search variable selection technique. We explore the performance of our proposed approach on both a synthetic dataset and several realistic document datasets. The comparison between our proposed approach and stage-of-the-art document clustering approaches indicates that our approach is robust and effective for document clustering.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications-Data Mining; I.5.3 [Pattern Recognition]: Clustering

General Terms

Algorithms

Keywords

Document Clustering, Dirichlet Process Mixture Model, Feature Selection.

1. INTRODUCTION

With the rapid growth of Internet and the wide availability of news documents, document clustering, as one of the most useful tasks in text mining, has received more and more interest recently. A common challenge in document clustering is to determine the number of document clusters *K*. This issue is not considered by most of the existing document clustering approaches [9, 18, 22]. They all take the assumption that *K* is a pre-defined parameter determined by users and provided before the document clustering process. However, given a set of documents, users need to browse the whole document collection in order to estimate *K*. This is obviously time consuming and unrealistic especially when the size of document collection is extremely large. Moreover, an improper estimation of *K* might easily mislead the clustering process and result in bad clustering outcome. Therefore, it is useful if a document clustering approach could be designed relaxing the assumption of the pre-defined *K*.

Determine the number of clusters is a difficult problem. We attempt to group documents into an optimal number of document clusters based on the Dirichlet process mixture (DPM) model. The DPM model has been studied in nonparametric Bayesian for a long time [1, 14, 21]. As an infinite mixture model in which each component corresponds to a different cluster, the DPM model determines the number of clusters automatically. When a new data point arrives, it either rises from existing clusters or starts a new cluster. The clustering process based on the DPM model jointly considers both the data likelihood and the clustering property of the DP prior that data points are more likely to be related to popular and large clusters [2, 10]. This flexibility of the DPM model makes it particularly useful for document clustering. However, there is no work investigating DPM model for document clustering. One reason is that the high-dimensional representation of text documents is composed of all distinct words including discriminative words and a large number of irrelevant noise words. In [17], it is mentioned that the optimal number of clusters is greatly affected by the quality of the feature subset. The involvement of irrelevant words confuses the process of estimating the optimal number of clusters *K* which causes poor clustering solution in return. Therefore, it is necessary to separate discriminative words from irrelevant noise words and only use them to group document collection especially when *K* is unknown.

In this paper, we propose an approach, namely Dirichlet process mixture model with feature selection (DPMFS), which 1) groups documents into a set of document clusters while *K* is determined automatically; 2) identifies discriminative words and separates them from irrelevant noise words. In our proposed approach, a DPM model is designed and investigated to group documents as well as discover the optimal number of document clusters. The DPM model is not without problems. One problem for DPM is that DPM parameters cannot be estimated quickly. In our proposed approach, a Dirichlet Multinomial Allocation (DMA)

KDD'10, July 25-28, 2010, Washington, DC, USA.
Copyright 2010 ACM 978-1-4503-0055-1/10/07...$10.00.

model is used to approximate the DPM model. To identify discriminative words, a stochastic search variable selection technique [5, 12, 16] is applied. In our inference procedure, the Gibbs sampling algorithm [14, 21] is used to infer both the cluster structure and the discriminative words. We have conducted extensive experiments on our proposed DPMFS approach by using both synthetic and realistic datasets. We also compared our approach with a stage-of-the-art model-based document clustering approach proposed in [9] and a standard model-based clustering approach [24]. Experimental results show that our proposed DPMFS approach is effective.

The remainder of this paper is organized as follows: First, related work on the identification of the number of clusters and document clustering is discussed in section 2. In section 3, we introduce background knowledge of the DPM model and the DMA model. Next, in section 4, we describe the DPMFS model and DMAFS model. Our proposed algorithm is given in section 5. Section 6 presents the design of experiments and discusses results of experiments. Finally, in section 7, we draw conclusions and make suggestions for future work.

2. RELATED WORK

Many methods have been introduced to find an optical number of clusters *K*. The most straightforward method is the likelihood cross-validation technique [27] which trains the model with different values of *K* and then picks the one with the highest likelihood on some held-out data. Another solution is to assign a prior to *K* and then calculate the posteriori distribution of *K* to determine this number [6]. In the literature, there are also many information criteria proposed to choose *K*, e.g., Minimum Description Length (MDL) [23], Minimum Message Length (MML) [30], Akaike Information Criterion (AIC) [4] and Bayesian Information Criterion (BIC) [25]. 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 trade-off data likelihood and model complexity [11]. Compared to all these methods, the method based on the DPM model to choose *K* is very different and flexible. In the DPM model, the number of clusters is determined after the clustering process rather than pre-estimated. Furthermore, this method is easy to use and does not require very expensive computation. In the previous work, [29] applies DPM model to the lexical-semantic verb clustering and [3] uses this model in the image analysis. They all pointed that DPM model could determine appropriate number of clusters automatically.

If the number of clusters is pre-defined, many algorithms based on the probabilistic finite mixture model have been successfully applied to the document clustering. For example, [22] proposed a multinomial mixture model. It applies the EM algorithm for document clustering assuming that document topics follow multinomial distribution and each document is a mixture of these multinomial distributions. This method has been shown to perform well for the document dataset though it does not take into account the phenomenon that words in a document tend to appear in bursts. [19] used the DCM model to capture burstiness well. Their 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. [9] derived the EDCM distribution which belongs to the exponential family and it is a good approximation to the DCM distribution. The EM algorithm with the EDCM distribution is faster than the corresponding algorithm with DCM distribution proposed in [19]. EM algorithm with EDCM distribution is the most competitive in the literature for document clustering in recent years.

3. BACKGROUD

3.1 Dirichlet Process Mixture Model

The DPM model is a mixture model with an infinite number of mixture components [28]. We introduce this infinite mixture model by firstly 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_n is drawn from one of *K* multinomial distributions parameterized by *K* different multinomial parameters, $\theta_1,\ldots,\theta_K$. Since the number of clusters is always unknown, to allow it to grow with data, we assume that the data point x_n follows a general mixture model in which the parameter θ is generated from a distribution *G*. The conditional hierarchical relationships are as follows:

$$\theta_n \mid \mathrm{G} \;\sim\; \mathrm{G},\; n = 1,2,\ldots,D,$$

$$x_n \mid \;\theta_n \sim F(\,x_n \mid \theta_n),\, n = 1,2,\ldots,D, \tag{1}$$

where D is the number of data points and $F(x_n|\theta_n)$ is the distribution of x_n given the parameter θ_n.

In the general mixture model, probability distribution *G* is always unknown. If the unknown *G* is a discrete distribution on a finite set of values, this general mixture model reduces to the finite mixture model. Bayesian nonparametric methods view *G* as a (infinite-dimensional) parameter and assign a prior to it. One class of Bayesian nonparametric techniques is called the Dirichlet process (DP) [10].

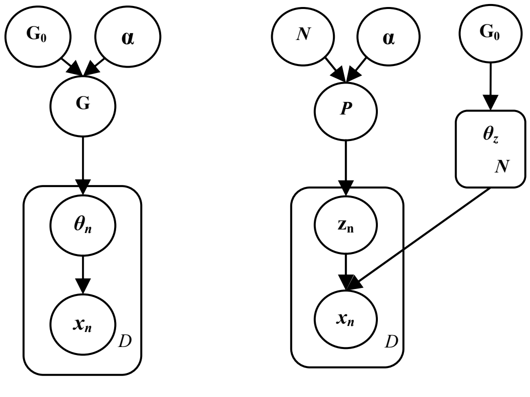

Figure 1: Graphical representation of DPM model (Left) and DMA (Right).

Dirichlet process, as a distribution on distributions, is parameterized by a positive scaling parameter α and a base distribution G_0. Assigning a DP prior to G in the general mixture model leads to the Dirichlet process mixture (DPM) [1] model. The hierarchical Bayesian specification of DPM model is as follows:

$$\mathrm{G} \mid \alpha, \mathrm{G}_0 \;\sim\; \mathrm{DP}\,(\alpha, \mathrm{G}_0),$$

$$\theta_n \mid G \sim G,\ n = 1,2,\ldots,D, \qquad (2)$$

$$x_n \mid \theta_n \sim F(x_n \mid \theta_n),\ n = 1,2,\ldots,D.$$

The DPM model can be best understood by the hierarchical graphical representation shown in Figure 1.

As shown in [1], integrating out G, the joint distribution of the collection of variables $\{\theta_1,\ldots,\theta_D\}$ exhibits a clustering effect. Let θ_{-n} denotes the set of all θ_j for $j\neq n$. The conditional distribution of θ_n given θ_{-n} has the following form:

$$\theta_n \mid \theta_{-n}, \alpha, G_0 \sim \frac{1}{D-1+\alpha}\sum_{j\neq n}\delta_{\theta_j} + \frac{1}{D-1+\alpha}G_0. \qquad (3)$$

Let $\Phi_1,\ldots,\Phi_C$ be the distinct values taken by θ_{-n} where C is the number of clusters estimated. Let m_i be the number of times that the value of θ_j equals to Φ_i for $j\neq n$. Equation (3) is transformed to:

$$\theta_n \mid \theta_{-n}, \alpha, G_0 \sim \sum_{i=1}^{C}\frac{m_i}{D-1+\alpha}\delta_{\Phi_i} + \frac{\alpha}{D-1+\alpha}G_0. \qquad (4)$$

Equation (4) means that parameters $\theta_1, \ldots,\theta_D$ are randomly partitioned into clusters, in which all θ take on the same value. It also indicates that DP prior allows a new data point either to share the same cluster with the previous data points or to start a new cluster. The number of clusters is determined automatically. We can best understand this clustering property by a famous metaphor known as the Chinese restaurant process [28].

Given data points $x_1,\ldots, x_D$ and the DP parameter (α, G_0), DPM model yields a posterior distribution on $\theta_1, \ldots,\theta_D$ which also exhibits clustering effect [21]. Based on the posterior estimation of $\theta_1, \ldots,\theta_D$, the data points $x_1,\ldots\ x_D$ can be partitioned into clusters. Data points in cluster i share the same parameter value Φ_i. Since this clustering process based on the DPM model not only considers the data likelihood as the finite mixture model but also combines the clustering property of the DP prior shown in Equation (4), the DPM model is very suitable for document clustering.

3.2 Dirichlet Multinomial Allocation

It has been proved 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 [13, 15, 20]. The Dirichlet Multinomial Allocation (DMA) [13] is one of the most famous approximations to the DPM model. The generative model for DMA is as follows:

$$x_n \mid z_n, \theta \sim F(\theta_{z_n}),\ n = 1,\ldots,D,$$

$$z_n \mid p \sim \mathit{Discrete}\ (p_1,\ldots,p_N),\ n=1,\ldots,D, \qquad (5)$$

$$\theta_z \sim G_0,$$

$$p \sim \mathit{Dirichlet}\ (\alpha/N,\ldots,\alpha/N),$$

where z_n indicates the latent cluster allocation of the n-th sample and N is the number of mixture components. For each cluster z, the parameter θ_z determines the distribution of the data points from that cluster. The N-dimensional vector p, which is the mixing proportions for the clusters, is given a Dirichlet prior with symmetric parameters α/N. The graphical representation of DMA is shown in Figure 1.

Let z_{-n} denote the set of all z_j for $j\neq n$. Integrating out the mixing proportions p, we can write the conditional distribution of z_n given z_{-n} as the following form:

$$p(z_n = z \mid z_{-n}) = \frac{n_{n,z} + \alpha/N}{n-1+\alpha}, \qquad (6)$$

where z ranges from 1 to N and $n_{n,z}$ is the number of z_j for $j\neq n$ that are equal to z.

Compare the Equation (4) and the Equation (6), the clustering property of the DMA is the same as DPM model if we let N go to infinity. It has been shown in [14] that the L_1 distance between the Bayesian marginal density of the data under DMA and the DPM model is $O(4D\ exp(-(N-1)/\alpha))$. This property provides good hints on how to choose the value of N. For example, if D=300, N=30, and α=1.0, we get an L_1 bound of 3.05E-10. Therefore, for D=300 and α=1.0, a DMA model with N=30 is virtually indistinguishable from the DPM model.

4. DPMFS AND DMAFS APPROXIMATION

Suppose there are D documents in a dataset x with the vocabulary size W. The set of vocabulary is composed of all words appeared in x represented as $\{w_1, w_2, \ldots, w_W\}$. Given a document x_i in X, let x_{ij} be the number of appearances of the word w_j. Each document is represented as a W-dimensional vector $x_i= (x_{i1}, x_{i2},\ldots, x_{iW})$.

4.1 Stochastic Search Variable Selection

We introduce a latent binary vector $\gamma=(\gamma_1,\ldots, \gamma_W)$ to identify words that discriminate between the different clusters.

$$\gamma_j = \begin{cases} 1, \text{ if } w_j \text{ is discriminative,} \\ 0, \text{ otherwise.} \qquad j = 1,\ldots,W. \end{cases} \qquad (7)$$

This latent vector partitions the dataset x into two parts: one part is the discriminative words, $x\gamma= \{(x_{i1}\gamma_1,\ldots, x_{iW}\gamma_W)$: i=1,2,…, $D\}$ which defines the latent cluster structure. Another part is the irrelevant noise words, $x(1-\gamma) =\{(x_{i1}(1-\gamma_1),\ldots, x_{iW}(1-\gamma_W))$: i=1,2,…, $D\}$ that confuses document clustering process. We assign a prior to γ and assume that its elements are independent Bernoulli random variables with common probability distribution. The distribution of γ is as follows:

$$p(\gamma) = \prod_{j=1}^{W}\omega^{\gamma_j}(1-\omega)^{1-\gamma_j}, \qquad (8)$$

where ω is the prior probability of each word expected to be discriminative.

This stochastic search variable selection technique has been used successfully in various applications to identify informative variables [12, 16]. As [16], we will combine this technique with DPM model and DMA in Section 4.2 - 4.3.

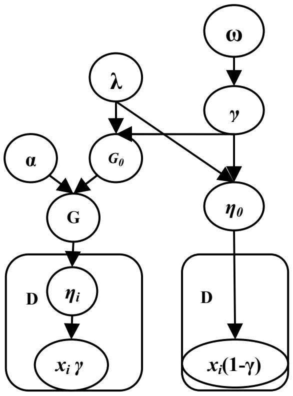

Figure 2: Graphical representation of DPMFS model.

4.2 DPM Model with Feature Selection

We assume the following generative process for the D documents in a dataset:

1. Choose $\gamma \mid \omega \sim p(\gamma)$.
2. Choose $N_{ij} \sim Poisson\ (\xi_j)$, $i = 1, 2,\ldots, D, j = 1,2$.
3. Choose G | γ, λ ~ DP (α, G_0), where $\lambda = (\lambda_1,\ldots,\lambda_W)$ and G_0 is a Dirichlet distribution with parameter $\lambda_1\gamma_1,..., \lambda_W\gamma_W$.
4. Choose η_i | G ~ G, $i = 1, 2,\ldots, D$.
5. Choose $\eta_0 \mid \gamma, \lambda \sim Dirichlet\ (\lambda_1(1\text{-}\gamma_1),\ldots, \lambda_W(1\text{-}\gamma_W))$.
6. Choose $x_i\gamma \mid \eta_i \sim Multinomial\ (\eta_i; N_{i1})$, $i = 1,\ldots, D$.
7. Choose $x_i(1\text{-}\gamma) \mid \eta_0 \sim Multinomial\ (\eta_0; N_{i2})$, $i = 1,\ldots, D$.

where $p(\gamma)$ is shown in Equation (8), N_{i1} is the total appearances of the discriminative words in document x_i and N_{i2} is the total appearance of the irrelevant noise words in x_i. N_{i1} and N_{i2} are both unobservable and considered as latent variable. $x_i\gamma$ and $x_i(1\text{-}\gamma)$ represent $(x_{i1}\gamma_1,\ldots,x_{iW}\gamma_W)$ and $(x_{i1}(1\text{-}\gamma_1),\ldots, x_{iW}(1\text{-}\gamma_W))$ respectively. η_i denotes the multinomial parameter for the discriminative words in x_i and η_0, as the multinomial parameter for the irrelevant noise words, is shared by all the documents in the dataset.

The graphical representation of DPMFS model is shown in Figure 2. From the generative process, it is not difficult to find that DPM model is only used to model the data with discriminative words, in particular, $x_i\gamma$, $i = 1,2,\ldots, D$. Parameters in the Dirichlet distribution and Multinomial distribution used in the our model may be zero. We only consider those non-zero parameters. For example, the probability density functions for $x_i\gamma$ is as follows:

$$f(x_i\gamma \mid \gamma, \eta_i) = \frac{N_{i1}!}{\prod_{\substack{j=1\\ \gamma_j=1}}^{W} x_{ij}!} \prod_{\substack{j=1\\ \gamma_j=1}}^{W} \eta_{ij}^{x_{ij}} . \tag{9}$$

In our model, words in each document are divided into two parts according to whether they define the underlying cluster structure. We assume that there is no correlation between the set of discriminative words and the set of irrelevant noise words. So the probability density function for x_i is given by:

$$f(x_i \mid \gamma, \eta_i, \eta_0) = f(x_i\gamma \mid \eta_i) f(x_i(1\text{-}\gamma) \mid \eta_0). \tag{10}$$

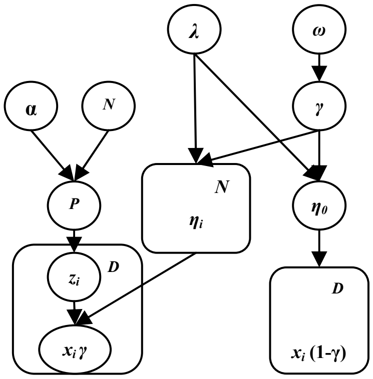

Figure 3: Graphical representation of DMAFS model.

4.3 Approximating the DPMFS Model

In this section, we design a DMA model with feature selection, named DMAFS. Since the DPM model can be approximated by the DMA, it is obvious that the DMAFS model is also a good approximation to the DPMFS model. The DMAFS assumes the following generative process for each document x_i in a dataset:

1. Choose $\gamma \mid \omega \sim p(\gamma)$.
2. Choose $N_{ij} \sim Poisson\ (\xi_j)$, $i = 1, 2,\ldots, D, j = 1,2$.
3. Choose $\eta_i \mid \gamma, \lambda \sim Dirichlet\ (\lambda_1\gamma_1,\ldots, \lambda_W\gamma_W)$, $i = 1,..., N$.
4. Choose $\eta_0 \mid \gamma, \lambda \sim Dirichlet\ (\lambda_1(1\text{-}\gamma_1),\ldots, \lambda_W(1\text{-}\gamma_W))$.
5. Choose $p \mid \alpha \sim Dirichlet\ (\alpha/N,\ldots, \alpha/N)$.
6. Choose $z_i \mid p \sim Discrete\ (p_1,\ldots, p_N)$, $i = 1, 2,..., D$.
7. Choose $x_i\gamma \mid \eta_1,...,\eta_N, z_i \sim Multinomid(\eta_{z_i}; N_{i1})$, $i = 1,\ldots, D$.
8. Choose $x_i(1\text{-}\gamma) \mid \eta_0 \sim Multinomial\ (\eta_0; N_{i2})$, $i = 1,\ldots, D$.

A graphical representation of DMAFS model we proposed is shown in Figure 3. The DMAFS approximation provides a close connection between finite mixture model and infinite mixture model. It allows us to have a better understanding of the data generative process from DPMFS model by comparing the finite mixture model. Furthermore, the DMAFS model is very useful to derive simple and effective Gibbs sampling algorithm for DPMFS model. The Gibbs sampling algorithm is shown in Section 5.

Since Dirichlet distribution is the conjugate prior for the parameter of multinomial distribution, integrating over $\eta_0, \eta_1,\ldots, \eta_N$ in Equation (10), the likelihood of the D documents conditioned on the latent variables γ and z becomes:

$$f(x \mid \gamma, z) = (\prod_{i=1,D} T_{i(\gamma)}) \cdot S_{1(\gamma)} \cdot S_{2(\gamma)} \cdot Q_{(\gamma)}^{M} \prod_{k=1,N} R_{k(\gamma)}, \tag{11}$$

in which M is the number of distinct values taken by z and

$$T_{i(\gamma)} = \frac{(\sum_{j=1,W} x_{ij}\gamma_j)!\,(\sum_{j=1,W} x_{ij}(1-\gamma_j))!}{\prod_{j=1,W} x_{ij}!},$$

$$S_{1(\gamma)} = \frac{\Gamma(\sum_{j=1,W} \lambda_j(1-\gamma_j))}{\Gamma(\sum_{i=1,D}\sum_{j=1,W} x_{ij}(1-\gamma_j) + \sum_{j=1,W} \lambda_j(1-\gamma_j))},$$

$$S_{2(\gamma)} = \prod_{\substack{j=1,W \\ \gamma_j=0}} \frac{\Gamma(\sum_{i=1,D} x_{ij} + \lambda_j)}{\Gamma(\lambda_j)}, \quad Q_{(\gamma)} = \frac{\Gamma(\sum_{j=1,W} \lambda_j \gamma_j)}{\prod_{\substack{j=1,W \\ \gamma_j=1}} \Gamma(\lambda_j)},$$

$$R_{k(\gamma)} = \frac{\prod_{\substack{j=1,W \\ \gamma_j=1}} \Gamma(\sum_{\{i:z_i=k\}} x_{ij} + \lambda_j)}{\Gamma(\sum_{\{i:z_i=k\}} \sum_{j=1,W} x_{ij}\gamma_j + \sum_{j=1,W} \lambda_j \gamma_j)}.$$

5. ALGORITHM

We use the Gibbs sampling method to infer both the latent cluster structure and discriminative words in the context of DMAFS model. The inference procedure is effective for the DPMFS model if we choose the parameter N large enough following the advice of [14].

Let the state of Markov chain consist of $\gamma= \{\gamma_1,\ldots,\gamma_W\}$, $\eta=\{\eta_0, \eta_1,\ldots,\eta_N\}$ and $z=\{z_1,\ldots,z_D\}$. Let $\{z_1^*,\ldots,z_M^*\}$denote the set of distinct values of z. Our inference procedure is as follows:

1. Initialize the latent variables γ and z, set the parameter α, ω, λ and N.
2. Update the latent discriminative words indicator γ by repeating the following Metropolis step R_1 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

$$\min\{1, \frac{f(\gamma_{new} | x, z)}{f(\gamma_{old} | x, z)}\}, \tag{12}$$

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

3. Conditioned on the other latent variables, for k =1,…, N, if k is not in $\{z_1,\ldots,z_D\}$, update η_k by sampling a value from a Dirichlet distribution with parameter $\lambda_1\gamma_1,\ldots, \lambda_W\gamma_W$. For i =1,…, M, update $\eta_{z_i^*}$ by sampling a value from a Dirichlet distribution with the following parameters:

$$\sum_{\{j:z_j=z_i^*\}} x_{jl}\gamma_l + \lambda_l\gamma_l,\ l = 1,\ldots, W. \tag{13}$$

4. For i =1,2,…, D, update the latent data label z_i by repeating the following Metropolis step R_2 times: A new candidate z_i^{new} is drawn from the following distribution:

$$p(z_i^{new} = z | z_{-i}) = \frac{n_{iz} + \alpha/N}{D-1+\alpha}. \tag{14}$$

   where z_{-i} denotes all the z_j for $j\neq i$ and n_{iz} is the number of z_j for $j\neq i$ that are equal to z. This new candidate is accepted with the probability:

$$\min\{1, \frac{f(x_i\gamma | \eta_{z_i^{new}})}{f(x_i\gamma | \eta_{z_i})}\}. \tag{15}$$

5. Update λ if necessary by the following sampling:

   5a. update η_0 by sampling a value from a Dirichlet distribution with the following parameters:

$$(1-\gamma_l)(\sum_{i=1,D} x_{il} + \lambda_l),\ l = 1,\ldots W. \tag{16}$$

   5b. Assign a prior $p(\lambda)$ to λ and draw λ from

$$p(\lambda | \gamma, \eta_0, \eta_1 \ldots \eta_N) \propto p(\lambda) p(\eta_0 | \lambda, \gamma) \prod_{i=1,N} p(\eta_i | \lambda, \gamma). \tag{17}$$

6. After sampling γ, η, z and λ by step 2-5 for many times (known as “burn-in” period), we use the last H samples of z and γ to infer the latent data label and discriminative words as follows:

   6a. The estimated label of document x_i is the most frequent value of z_i in the last H samples.

   6b. The jth word is discriminative if the average value of the last H samples of γ_j is bigger than a threshold such as 0.7 which is used in our experiments.

Note that our inference procedure only focuses on the latent variables γ, η and z which are closely related with the cluster structure or the discriminative word subset. The other latent variables such as p are integrated out. We use a simple initialization method to initialize γ and z. The initial label of each document is selected randomly from 1, 2,…, N. We randomly choose one discriminative word from those words appearing in the dataset. Because η is sampled in step 3, we don’t have to initialize it. The advice for choosing the parameters is discussed in Section 6.1.3.

6. EXPERIMENTS

We describe two sets of experiments to evaluate the performance of the DPMFS approach. For the first set of experiments, a synthetic dataset is used. For the second set of experiments, the DPMFS approach is evaluated using a set of real document datasets.

6.1 Evaluation Metric

We used the normalized mutual information (*NMI*) [8] 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 practice, *NMI* is estimated as follows [26]:

$$NMI = \frac{\sum_{h,l} d_{h,l} \log(\frac{d \cdot d_{h,l}}{d_h c_l})}{\sqrt{(\sum_h d_h \log(\frac{d_h}{d}))(\sum_l c_l \log(\frac{c_l}{d}))}} \tag{18}$$

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 $d_{h,l}$ 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 Dataset

6.2.1 *Dataset and Experimental Setup*

We have generated a synthetic dataset for conducting experiments. The synthetic data consisted of 300 data points with 1000 features.

Data points were generated by two different processes with four multinomial distributions. The first process was used to generate discriminative features. Specially, the first 50 features were regarded as discriminative features generated from a multinomial mixture distribution with three components. Each component represents one cluster and each cluster contains 100 data points. The second process was used to generate the irrelevant noise features. In particular, the remaining 950 features were regarded as irrelevant noise features generated from a multinomial distribution. The data was generated as follows:

$(x_{i1},\ldots,x_{i50}) \sim \textit{Multinomial}\ (\pi_j;\ 100),\ i=1+100(j\text{-}1),\ldots,100j,\ j=1,2,3.$

$(x_{i51},\ldots,x_{i1000}) \sim \textit{Multinomial}\ (\pi^*;\ 100),\ i=1,\ldots,\ 300.$

where $(\pi_1;\ 100)$, $(\pi_2;\ 100)$, $(\pi_3;\ 100)$ and $(\pi^*;\ 100)$ are the multinomial parameters. π_1, π_2, π_3 and π^* are chosen randomly in our experiment.

In our proposed algorithm for this synthetic data, we set N=30, R_1 =R_2 =5, α =1.0, ω=0.01. The components of parameter λ were all chosen to be 0.1. We ran our proposed algorithm 30 times and each time we ran 2500 iterations in which the first 2000 as burn-in.

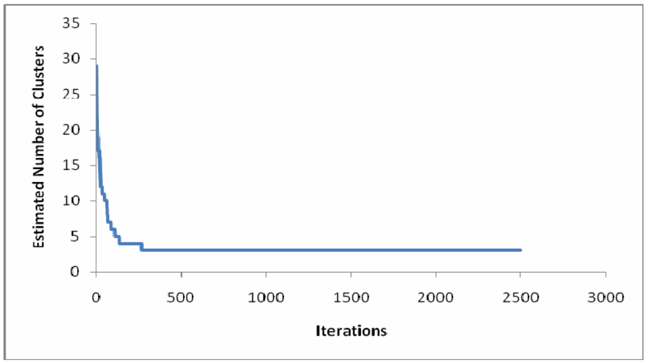

Figure 4: Trace plot for the number of clusters.

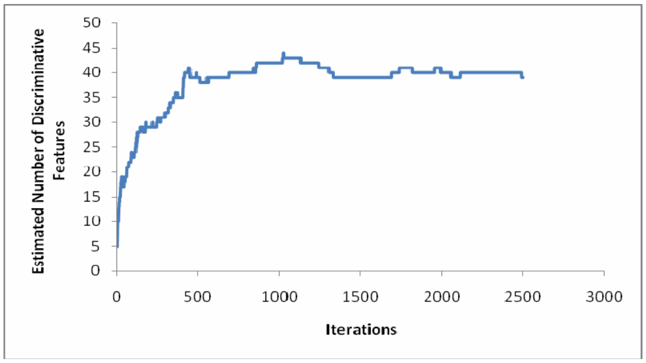

Figure 5: Trace plot for the number of discriminating features.

6.2.2 Experimental Performance

Our algorithm identified the perfect cluster structure for all the 30 runs of experiments. The number of features identified as discriminative stabilized around 40 to 45. On average, there were 41 true discriminating features identified successfully. Figure 4 and Figure 5 depict the number of clusters and the number of discriminative features estimated in one typical run by varying the number of iterations. The result shows that the number of clusters is faster to stabilize than the feature selection process. Moreover, our experiment indicated that the document clustering assignments also stabilized long before the number of discriminative features reach to a stable value. One possible reason is that the documents could be grouped with a subset of discriminative words. Since the purpose for document clustering is to group the dataset into an optimal partition, the sampling process could be terminated when the cluster assignment is not changed.

6.2.3 Discussion

We investigated the sensitivity of the choices of parameters in our algorithm by large amounts of experiments.

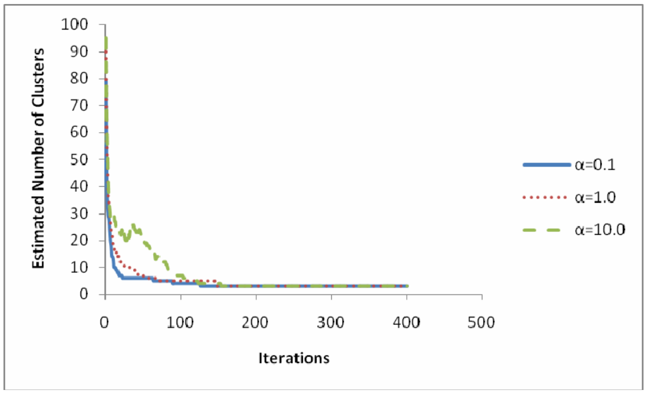

Figure 6: Trace plot for the number of clusters when α is chosen to be different values (Only show the first 400 iterations).

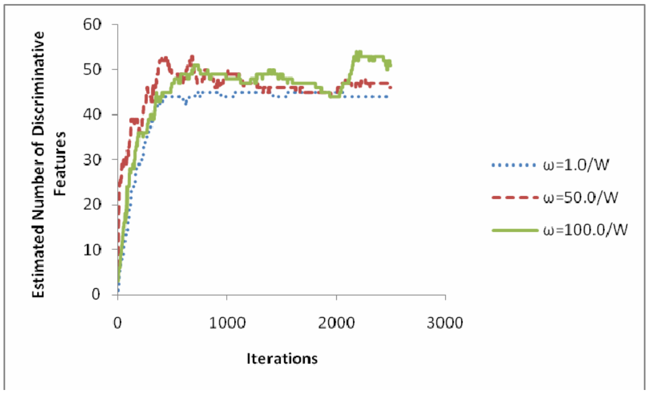

Figure 7: Trace plot for the number of discriminating features when ω is chosen to be different values.

Choice of N, R_1 and R_2: In principle, we can choose N to be the number of data points. However, in order to save computing time, we could choose a relatively small N follow the advice of [14] as mentioned in Section 2. Normally, in order to keep the DMA approximate the DPM model well, we should set a large N if we have chosen a large α. The number of Metropolis step R_1 and R_2 were both chosen to be 5 in our algorithm because we found that a bigger value had little improvement in the clustering quality though it would make the feature selection process more stable.

Choice of α and ω: We investigated the sensitivity of the choice of parameters α and ω which influenced the estimated number of clusters and the estimated number of discriminative features respectively. We simulated with different values of α where α was set to be 0.1, 1.0 and 10.0 which corresponds to a small, moderate, large prior number of clusters in the data under the DPM model. We also experimented with different values of ω where ω was set to be a small value 1.0/*W*, a moderate value 50.0/*W*, and a very high value 100.0/*W*. For the three different values of α, ω was fixed as 0.01 and *N*=200. For the three different values of ω, α was fixed as 1.0 and *N*=30. The other parameters were chosen to be: R_1 =R_2 =5 and the component of λ = 0.5. Our proposed approach achieved perfect clustering structure in all these experiments. This indicates that our algorithm is robust to the choice of α and ω. Figure 6 and Figure 7 show the trace plot of the estimated number of clusters and the estimated number of discriminative features respectively. Figure 6 indicates that a large α requires relatively long time for the estimated number of clusters to be stable. This is because a large value of α will make the model generate a new cluster easily as shown in Equation (4). From Figure 7, we found that a large value of ω made the sampler visit models with more discriminative features. However, the final estimated discriminative features by the last 500 samples were almost the same for all these different ω.

Choice of λ: The parameter λ not only affects the estimated number of clusters but also the estimated number of discriminating features. Some care is needed to choose this parameter in a reasonable range since a much larger value for it will result in a model with fewer mixture components than the true one. Our experiments indicate that a small value for λ performs well though it will require relatively long time for the sampling process to be stable. In order to acquire good clustering quality and save computing time, we must consider the characteristic of the dataset for setting the value of λ. For the document datasets used in our following experiments, we found that a good choice of λ_j is 1.0/σ_j, where σ_j is the sample standard variance of {x_{1i}, x_{2i},..., x_{Di}}. A simple interpretation for this choice is that if the standard variance of each column of the data *x* is large, there may be more clusters in the corpus and we should choose a small λ. Another effective method to handle the parameter λ is to place a prior to it and update this parameter in the inference procedure. This will require additional slow Metropolis-Hastings updates in our algorithm because we don't know the conjugate prior for this Dirichlet parameter λ.

Table 1: Datasets Description

(*D*: Number of documents, *K*: Number of clusters, *W*: Vocabulary size.)

Datasets	*D*	*K*	*W*
News-Different-3	300	3	2121
News-Similar-3	300	3	1767
News-Moderated-6	600	6	4036
Classic400	400	3	6025

6.3 Real Document Datasets

6.3.1 Experimental Datasets

Four standard text datasets were used in our experiments: *News-Different-3*, *News-Similar-3*, *News-Moderated-6* and *Classic400*. The summary of these four real-world text document datasets is shown in Table 1. The first three datasets were derived from the *20-Newsgroups* collection. This collection has messages collected from 20 different Usenet news-groups, 1000 messages from each newsgroup. From the original corpus, a subset was first created by randomly selecting 100 messages from each of the 20 newsgroups. The first three datasets were then derived from the subset. *News-Different-3* consists of 300 messages from 3 newsgroups on different topics (alt.atheism, rec.sport.baseball, sci.space) with well-separated clusters. *News-Similar-3* consists of 300 messages from 3 newsgroups on similar topics (comp.graphics, comp.os.ms-windows, comp.windows.x) where cross-posting often occurs. *News-Moderated-6* consists of 600 messages from 6 newsgroups on topics (rec.sport.baseball, sci.space, alt.atheism, talk.politics.guns, comp.windows.x, soc.religion.christian). In the *News-Moderated-6* dataset, some topics are similar (alt.atheism and soc.religion.christian) where others are different from each other. The *Classic 400* dataset, which is a typical unbalanced dataset, is the same dataset used by the EDCM model proposed in [9].

We pre-processed all the datasets by stop-word removal. Low frequency words were removed following the methodology presented in [7]. The purpose of such processing is to eliminate those words which obviously do not define the latent cluster structure. The threshold for removing low-frequency words for all datasets was set to 1.

6.3.2 Experimental Setup

For all the real world datasets experiments, we used the same setting of the parameters. The parameters were set as *N*=*D*/10, α=1.0, ω=50.0/*W*, R_1 =R_2 =5 and λ_j =1.0/σ_j, where *j* =1, 2,…, *W*. The initialization method for *γ* and *z* was the same as previous discussion in Section 4. Each time we ran 3000 iterations and the first 2500 as burn-in.

Table 2: Clustering results on *News-Similar-3* and *News-Moderated-6*

(*C*: Estimated number of clusters. EDCM and EM-MN use the true number of clusters).

Datasets	DPMFS		EDCM	EM-MN
	C	*NMI*	*NMI*	*NMI*
News-Similar-3	8.1	0.231	0.163	0.081
News-Moderated-6	7.9	0.663	0.531	0.562

For comparative investigation, a standard model-based clustering approach [22], labeled as EM-MN, was investigated as benchmark. We also ran experiments for a stage-of-the-art model-based clustering approach [9], labeled as EDCM. Since the deterministic annealing procedure [24] allows EM to find better local optima of the likelihood function and therefore improve the clustering quality, we added it to the EM-MN and EDCM. The temperature parameter for the three phases was chosen to be 25, 5 and 1. EM-MN and EDCM require the number of clusters as input. We studied the performance of them when we gave right or wrong number of clusters. Each algorithm was run 30 times and we used the average *NMI* to compare their performance for clustering.

Table 3: Clustering results on *News-Different-3* (the third row) and *Classic400* (the fourth row)

(*C*: Estimated number of clusters, *K*: Pre-defined number of clusters).

DPMFS		EDCM			EM-MN		
C	*NMI*	*K*=2	*K*=3	*K*=10	*K*=2	*K*=3	*K*=10
5.9	0.688	0.386	0.734	0.561	0.464	0.867	0.634
8.0	0.641	0.243	0.684	0.403	0.36	0.496	0.506

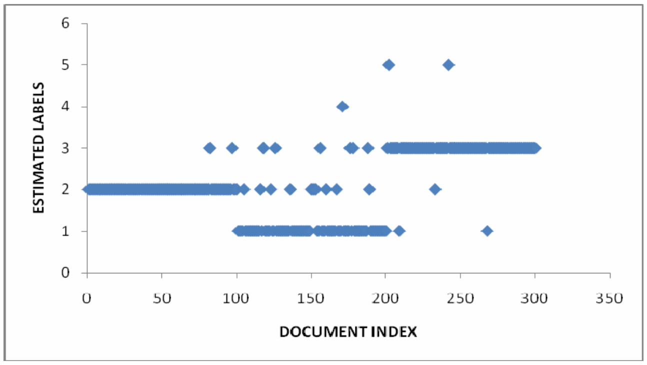

Figure 8: Estimated labels of data points in *News-Different-3*.

6.3.3 Experimental Results

Table 2 shows the experimental performances of DPMFS, EDCM and EM-MN on the *News-Similar-3* and *News-Moderated-6* datasets. The number of clusters estimated is also depicted. The experimental results show that our proposed approach achieves the best clustering results for these two datasets. The reason is that the *New-Similar-3* and the *New-Moderated-6* datasets contain similar clusters. A large number of irrelevant noise words in these two datasets may mislead the clustering process. This result demonstrates that DPMFS approach could separate the discriminative words from the irrelevant ones and therefore improve the clustering quality to some extent.

In Table 3, the experimental results on *News-Different-3* and *Classic400* datasets are depicted. The documents in these two datasets are relatively well separated and there are many discriminative words aiding the clustering process. It is shown in the experimental results that relatively better clustering quality were achieved by the EDCM and the EM-MN approaches when the number of clusters was correctly assigned. However, when the number of clusters was given imprecise, both of the EDCM and EM-MN approaches performed far worse than the DPMFS approach. Therefore, our proposed approach is more robust to group documents into a set of clusters when the number of clusters is unknown in advance.

In respect to the estimation of the number of clusters, the estimated values for these four datasets were all bigger than the true one as shown in Table 2 and Table 3. In fact, it is very difficult to acquire exact estimation of the number of document clusters in these datasets since a couple of outliers could make the estimated value bigger than the true one. Figure 8 shows one estimated labels of documents in *News-Different-3*. The estimated labels of the documents provide strong support for the true cluster structure. The result indicates that DPMFS could acquire meaningful clustering outcome.

7. CONCLUSIONS AND FUTURE WORK

In this paper, our proposed DPMFS approach handles document clustering and feature selection simultaneously. We constrain the DPM model only to define the cluster structure of the data with discriminative features which are identified by a latent binary vector. The Gibbs Sampling technique is used to infer both the cluster structure and the latent discriminative word subset. Our experiment shows that DPMFS approach groups document dataset into meaningful clusters without requiring the number of clusters known in advance. The comparison of our algorithm with some existing stage-of-the-art algorithms indicates that our approach is more robust and effective for document clustering when no information other than the observed values is available. Our analysis of the experiment result also shows that feature selection inserted in the DPM model could alleviate the negative impact of the irrelevant noise words and therefore improve the clustering quality.

An interesting direction for future research is to study how to use the DPMFS approach in the semi-supervised document clustering since more and more labeled documents or constraints are available in real-life. We think that the additional information could improve the clustering quality from at least two aspects. The first one is that reasonable model parameters and initial value can be chosen from this additional information. The second one is that we can use this information guide our sampling process.

8. ACKNOWLEDGMENTS

The work described in this paper was fully supported by grants from the Hong Kong Polytechnic University Internal Competitive Research Grants (Project Code: G-YG39 and A-PJ72).

9. REFERENCES

[1] C. Antoniak. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152-1174.

[2] D. Blackwell and J. MacQueen. (1973). Ferguson distribution via Polya urn schemes. *The Annals of Statistics*, 1(2):353-355.

[3] D. Blei and M. Jordan. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121-144.

[4] H. Bozdogan. (1983). Determining the number of component clusters in the standard multivariate normal mixture model using model-selection criteria. TR UIC/DQM/A83-1, Quantitative Methods Department, University of Illinois, Chicago, IL.

[5] P. J. Brown, M. Vannucci and T. Fearn. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, 60:627-641.

[6] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freedman. (1988). Autoclass: A Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 54-64.

[7] I. S. Dhillon and D. S. Modha. (2001). Concept decompositions for large sparse text data using clustering. *Journal of Machine Learning*, 42(1):143-175.

[8] B. E. Dom. (2001). An information-theoretic external cluster-validity measure. *Research Report RJ 10219,* IBM.

[9] C. Elkan. (2006). Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution. In *Proceedings of the 23th International Conference on Machine Learning*, 289-296.

[10] T. Ferguson. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209-230.

[11] C. Fraley and A. E. Raftery. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578-588.

[12] E. I. George and R. E. McCulloch. (1992). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881-889.

[13] P. J. Green and S. M. Richardson. (2001). Modelling Heterogeneity with and without the Dirichlet Process. *Scandinavian Journal of Statistics*, 28:355-377.

[14] J. Ishwaran and L. James. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161-174.

[15] H. Ishwaran and M. Zarepour. (2002). Exact and Approximate Sum-Representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269-283.

[16] S. Kim. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877-893.

[17] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(9):1154-1166.

[18] J. MacQueen. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

[19] R. Madsen, D. Kauchak, and C. Elkan. (2005). Modeling word burstness using the Dirichlet distribution. In *Proceedings of the 22th International Conference on Machine Learning*, 545-552.

[20] R. Neal. (1992). Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, 11:197-211.

[21] R. Neal. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249-265.

[22] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchel. (2000). Text classification from labeled and unlabeled documents using EM. *Journal of Machine Learning,* 39(2/3):103-134.

[23] J. Rissanen. (1978). Modeling by shortest data description. *Automatica*, 14:465-471.

[24] K. Rose. (1998). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. In *Proceedings of the IEEE*, 86(11):2210-2239.

[25] G. Schwarz. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461-464.

[26] Z. Shi. (2006). Semi-supervised model-based document clustering: A comparative study. *Journal of Machine Learning*, 65(1):3-29.

[27] P. Smyth. (1998). Model selection for probabilistic clustering using cross-validated likelihood. *ICS Tech Report 98-09, Statistics and Computing*.

[28] Y. W. Teh, M. I. Jordan, M.J. Beal, and D.M. Blei. (2007). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566-1581.

[29] A. Vlachos, Z. Ghahramani, and A. Korhonen. (2008). Dirichlet process mixture models for verb clustering. *ICML Workshop on Prior Knowledge for Text and Language Processing*, Helsinki, Finland.

[30] C. Wallace and P. Freedman. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B*, 49(3):240-265.