IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 5, July 2014.
www.ijiset.com
ISSN 2348 – 7968

Dirichlet Process Mixture Model For Document Clustering with Feature Extraction Which Helps In Page Ranking

[1]Nitesh Timande, [2]Dr. M.B.Chandak

M.Tech Scholar, WCEM, Nagpur
HOD, Dept of CSE, RKNEC, Nagpur

ABSTRACT

To figure out the appropriate number of clusters to which documents should be partitioned is crucial task in document clustering. In this paper, we propose a novel approach, namely DPMFS, which does document clustering and it helps in improving page ranking. Here we are firstly grouping documents into a set of clusters whereas the number of document clusters is determined automatically by the Dirichlet process mixture model secondly identifying the discriminative words and separate them from irrelevant noise words. Our experiment shows that our proposed approach performs well on the synthetic dataset as well as real datasets. The comparison between our approach and some other stage of the art document clustering approach shows that our approach is more robust as well as effective for document clustering.

Keywords
Database management, Dirichlet process mixture model, Document clustering, Feature Extraction.

1. INTRODUCTION

As the internet is growing rapidly 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. Document clustering, is nothing but grouping unlabeled text documents into meaningful clusters. Firstly, given a set of documents, users have to browse the whole document collection in order to find out *K* .This is not only time consuming but also unrealistic especially when dealing with large data sets

Many approaches of document clustering [9,18,22].Which were stated earlier they had not consider value of K .Furthermore, an improper estimation of *K* might easily mislead the clustering process. 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. They all take the assumption that *K* is a pre-defined parameter determined by users and provided before the document clustering process. We attempt for grouping documents into an optimal number of document clusters based on the Dirichlet process mixture (DPM) model.The DPM model figure out the number of clusters automatically.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.Two methods, variational inference and Gibbs sampling, are developed.

In this paper, we propose an approach, namely Dirichlet process mixture model with feature selection (DPMFS), which firstly does work of grouping documents into a set of document clusters where *K* is determined automatically and secondly identifing discriminative words and distribute them from irrelevant noise words. In our proposed approach, a DPM model is designed and investigated to group documents as well as to discover the optimal number of document clusters. To identify discriminative words, a stochastic search variable Selection technique [4, 9, 13] is applied.In our proposed approach , the Gibbs sampling algorithm [11] is used for inferring both the cluster structure as well as the discriminative words.

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 find the posteriori distribution of K to figure out this number [5]. 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].

In the DPM model, the number of clusters is determined after the clustering process rather than pre-estimated, this method is very easy to use and it aint require expensive computation. 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. Here we implemented the EM algorithm for document clustering assuming that document topics follow multinomial distribution and here each document is a combination of these multinomial distributions. This method has stated to perform well for the document dataset though it was not take into account the phenomenon that words in a document tend to appear in bursts, used the DCM model to capture burstiness well. Their work showed that the performance of DCM was comparable to that obtained with multiple heuristic changes to the multinomial model.

3. BACKGROUD

3.1Dirichlet Process Mixture Model

The DPM model is nothing but a mixture model with an infinite number of mixture components[28]. Here we will first describe 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 is used for document clustering assumes that each document xn is drawn from one of K multinomial distributions parameterized by K different multinomial parameters, θ1,.,θK. The conditional hierarchical relationships are as follows:

θn| G ~ G, n =1,2,…, D,
xn| θn ~ F(xn| θn), n =1,2,…, D, (1)

where D is the number of data points and F(xn|θn) is the distribution of xn given the parameter θn. In the general mixture model, probability distribution G is always unknown.One class of Bayesian nonparametric techniques is called the Dirichlet process (DP) [8].

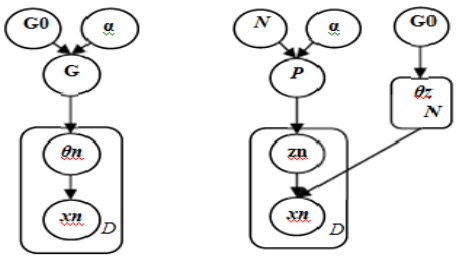

Figure 1: Graphical representation of DPM model (Left) and DMA (Right).

Dirichlet process, as a distribution on distributions, is parameterized by a positive scaling parameter α and a base distribution G0. The hierarchical Bayesian specification of DPM model is as follows:

$$G \mid \alpha, G0 \sim DP(\alpha, G0),$$
$$\theta_n \mid G \sim G, \; n=1,2,\ldots, D, \qquad (2)$$
$$x_n \mid \theta_n \sim F(x_n \mid \theta_n), \; n=1,2,\ldots, D.$$

The DPM model can be understood by the hierarchical graphical representation shown in Figure1,integrating out G, the joint distribution of the collection of variables {*θ1,…,θD*} exhibits a clustering effect. The conditional distribution of *θn* given *θ-n* has the following form:

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

Let Φ1 ,…,ΦC be the distinct values taken by θ-n where C is the number of clusters estimated. Let mi be the number of times that the value of θj equals to Φi for j≠n. Equation (3) is transformed to:

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

Equation (4) means that parameters θ1, .,θD are randomly partitioned into clusters, in which all θ take on the same value. The number of clusters is figured out automatically.Given data points x1,. ,xD and the DP parameter (α, G0), DPM model yields a posterior distribution on θ1, …,θDwhich also exhibits clustering effect. Based on the posterior estimation of θ1, …,θD, the data points x1,… xD can be partitioned into clusters. Data points in cluster share the same parameter value Φi.

3.2 Dirichlet Multinomial Allocation

It has been stated 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 [10, 12]. The Dirichlet Multinomial Allocation (DMA) [11] is one of the most famous approximations to the DPM model. The generative model for DMA is as follows:

$$x_n \mid z_n, \theta \sim F(\theta_{z_n}), n=1,\ldots, D,$$
$$z_n \mid p \sim Discrete(p_1,\ldots, p_N), \; n=1,\ldots, D, \qquad (5)$$
$$\theta_z \sim G_0,$$
$$p \sim Dirichlet(\alpha/N,\ldots, \alpha/N),$$

where zn indicates the latent cluster allocation of the n-th sample and N is the number of mixture components.Let z-n denote the set of all zj for j≠n. Integrating out the mixing proportions p, we can write the conditional distribution of zn 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 nn,z is the number of zj for j≠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 [11] that the L1distance between the Bayesian marginal density of the data under DMA and the DPM model is O(4D exp(-(N-1)/α)).

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 {w1, w2, …, wW}.Given a document xi in x, let

x, let xij be the number of appearances of the word wj. Each document is represented asW-dimensional vector xi= (xi1, xi,.,xiW)

4.1 Stochastic Search Variable Selection

We introduce a latent binary vector γ=(γ1 ,…, γW) to identify words that discriminate between the different clusters.

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

This latent vector partitions the dataset x into two parts: first part is the discriminative words, xγ= {(xi1γ1,…, xiWγW): i=1,2,.D} which defines the latent cluster structure. Another part is the irrelevant noise words, x(1-γ) ={(xi1(1-γ1),…, xiW(1-γW)): i=1,2,…,D} that confuses document clustering process. The distribution of γ is as follows:

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

where ω is the prior probability of each word expected to be discriminative.This stochastic search variable selection echnique has been used successfully in various applications to identify informative

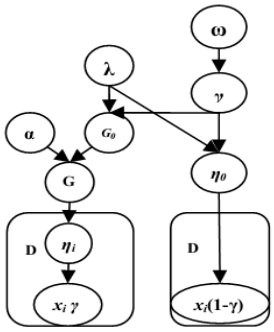

Figure 2: Graphical representation of DPMFS model.

4.2 DPM Model with Feature Selection

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

1. Choose γ | ω~ p(γ).
2. Choose Nij~Poisson (ξj),i =1, 2,…, D, j=1,2.
3. Choose G | γ, λ~ DP (α, G0), where λ= λ1,,λW) and G0 is a Dirichlet distribution with parameter λ1γ1 ,.λWγW
4. Choose ηi| G ~ G,i = 1, 2,…, D.
5. Choose η0 | γ, λ ~ Dirichlet(λ1 (1-γ1),…, λW (1-γW)).
6. Choose xi γ| ηi~ Multinomial (ηi; Ni1), i=1,…,D.
7. Choose xi(1-γ) |η0~Multinomial (η0; Ni2), i=1,…, D.

where p(γ) is shown in Equation (8), Ni1 is the total appearances of the discriminative words in document xi and Ni2 is nothing but the total appearance of the irrelevant noise words in xi. Ni1 and Ni2 are both unobservable and considered as latent variable.Parameters in the Dirichlet distribution and Multinomial distribution used in the our model may be zero. . 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, *xiγ*, *i* = 1,2,…, *D*.

For example, the probability density functions for xi γ is as follows

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

In our model, words in each document are divided into two parts according to whether they define the underlying cluster structure.

So the probability density function for xi is given by:

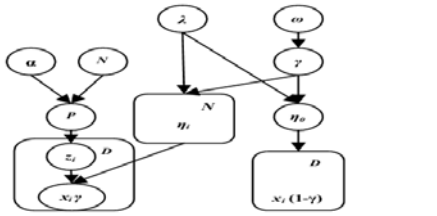

Figure 3: Graphical representation of DMAFS model.

4.3 Approximating the DPMF Model

In this section, we design a DMA model with feature selection, named DMAFS. The DMAFS assumes the following generative process for each document xi in a dataset:

1. Choose γ | ω~ p(γ).
2. Choose Nij~Poisson (ξj),i =1, 2,, D,j=1,2
3. Choose ηi |γ, λ Dirichlet(λ1γ1,..λWγW), i =1,..., N.
4. Choose η0 | γ, λ ~ Dirichlet(λ1 (1-γ1),…, λW (1-γW)).
5. Choosep | α~Dirichlet(α/N,…, α/N).
6. Choose zi| p ~ Discrete (p1…,pN), i=1,...,D
7. Choose , i=1,…, D.
8. Choose xi(1-γ) |η0~Multinomial (η0; Ni2), i=1,…, 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 variables[9,13] derive simple and effective Gibbs

ISSN 2348 – 7968

sampling algorithm for DPMFS model. The Gibbs sampling algorithm is shown in Section 5.The likelihood of the documents conditioned on the latent variables γ and z becomes:

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

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

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

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

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

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

5. ALGORITHM

Gibbs sampling method is used here to infer both the latent cluster Structure & discriminative words in the context of DMAFS model. Let the state of Markov chain consist of γ= {γ1,…,γW}, η={η0,η1,…,ηN} andz={z1,…,zD}. Let {z1*,…,zM*}denote the set of distinct values of z. Our inference procedure is as follows

1. Initialize the latent variables γand z, set the parameter α, ω, λand N.
2. Update the latent discriminative words indicator γ by repeating the following Metropolis step R1times: A new candidate γnew which adds or deletes a discriminative word is generated by randomly picking one of the W indices in γold and changing its value. The new candidate is accepted

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

3. Conditioned on the other latent variables, for k=1,…, N, if k is not in {z1,…,zD}, update ηk by sampling a value from a Dirichlet distribution with parameter λ1γ1,…, λWγW. For i =1,., M, update * η I z by sampling a value from a Dirichlet distribution with the following parameters:

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

4. For i=1,2,…, D, update the latent data label zi by repeating the following Metropolis step 2 times:

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

where z-I denotes all the zj for j≠I and niz is the number of zj for j≠i that are equal to z.

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

5. Update λ if necessary by the following sampling:

5a. update η0 by sampling a value from a Dirichlet distribution with the following parameters

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

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

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

6. After sampling γ, η, z and λ by step 2-5 for many times (known as “burn-in” period),

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

6b. The jth word is discriminative if the average value of the last H samples of γj is bigger than a threshold such as 0.7.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.

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]:

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

where d is the number of documents, dh is the number of documents in class h, cl is the number of documents in cluster l and dh,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 useful to generate the irrelevant noise features. The data was generated as follows:

($xi1$,…,$xi50$) ~ *Multinomial* (πj; 100), i=1+100(j-1),…,100j, j=1,2,3.

($xi51$,…,$xi1000$) ~ *Multinomial* (π^*; 100), i=1,…, 300.where ($\pi 1$; 100), ($\pi 2$; 100), ($\pi 3$; 100) and (π^*; 100) are the multinomial parameters. $\pi 1$, $\pi 2$, $\pi 3$ and π^* are chosen randomly in our experiment.

In our proposed algorithm for this synthetic data, we set N=30, $R1$=$R2$ =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.

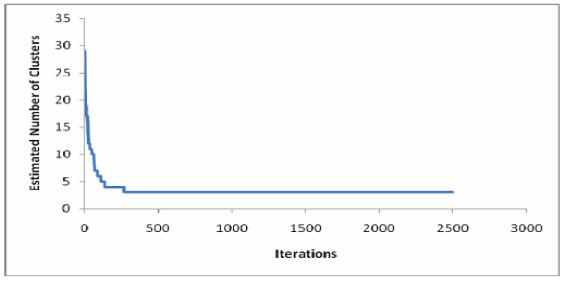

Figure 4: Trace plot for the number of clusters.

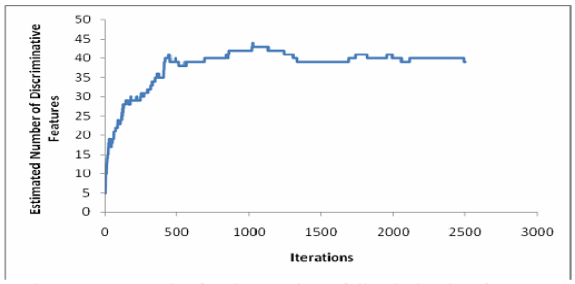

Figure 5: Trace plot for the number of discriminating features.

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.

6.2.3 Discussion

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

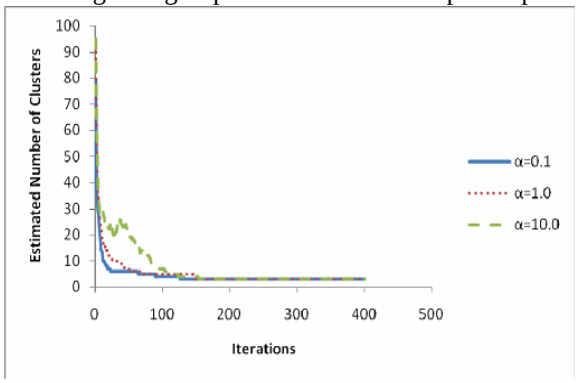

Figure 6: Trace plot for the number of clusters when α is chosen to be different values (Only show the first 400 iterations).

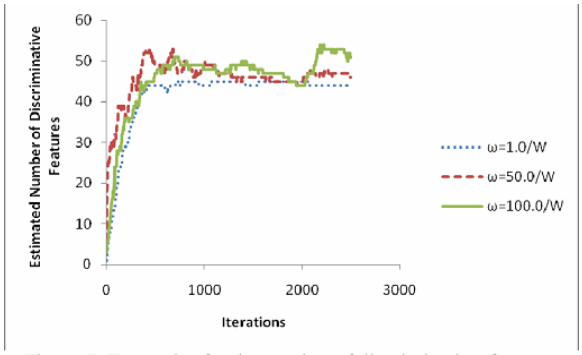

Figure 7: Trace plot for the number of discriminating features when ω is chosen to be different values.

Choice of N, $R1$ and $R2$: 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.

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.

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).

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.

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.atheismsoc.religion.christian) where others are different from each other. The *Classic 400* dataset, which is a typical unbalanceddataset, is the same dataset used by the EDCM model proposed in[9].

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*, *R1* =*R2* =5 and λj =1.0/*σj*, 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.

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

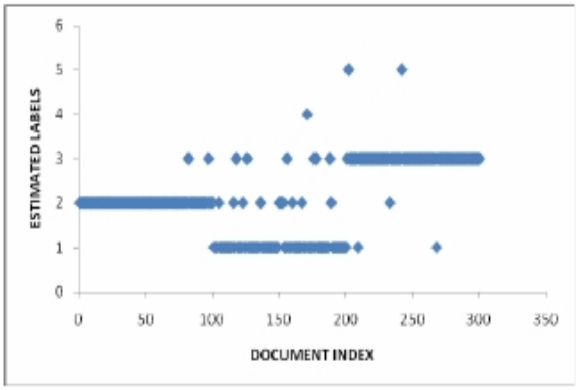

Figure 8: Estimated labels of data points in *News-Different-3*.

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 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 result indicates that DPMFS could acquire meaningful clustering outcome.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an approach which handles document clustering &feature extraction simultaneously which is helpful in improving page
ranking. Document words are partitioned according to their usefulness to discriminate the document clusters. The Gibbs Sampling technique is used to infer both the cluster structure and the latent discriminative word subset 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.

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. 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.

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 Polyaurn 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 FifthInternational 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 MachineIntelligence*, 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*,

ISSN 2348 – 7968

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 MachineLearning*, 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.

s