RESEARCH ARTICLE OPEN ACCESS

Document Clustering with Feature Selection Using Dirichlet Process Mixture Model and Dirichlet Multinomial Allocation Model

Nitesh Timande *, Dr. M.B.Chandak **, Prof.Manisha.Kamble***
**(M.Tech Scholar, Department of Computer Science, RTMN University, NAGPUR*
Email: nitesh.tim@gmail.com)
*** (HOD, Department of Computer Science, RTMN University, NAGPUR*
Email: chandakmb@rknec.com)
****(Asst.Prof. Department of Computer Science, RTMN University, NAGPUR*
Email: manisha.thool@gmail.com)

ABSTRACT
To find the appropriate number of clusters to which documents should be partitioned is crucial in document clustering. In this survey paper, we propose a novel approach, namely DPMFS, to address this issue. The proposed approach is designed firstly to group documents into a set of clusters while the number of document clusters is determined automatically by the Dirichlet process mixture model secondly to identify the discriminative words and separate them from irrelevant noise words via stochastic search variable selection technique. 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 paper indicate that our proposed approach performs well on the synthetic data set as well as real data sets.
Keywords– Bayesian Information Criterion, Dirichlet process mixture model, Document clustering, Feature Selection

I. 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. Document clustering, grouping unlabeled text documents into meaningful clusters in many application.

Firstly, 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 data sets 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]. Furthermore, an improper estimation of *K* might easily mislead the clustering process. If a bigger or a smaller number of clusters is used it ultimately degrades clustering accuracy .In this paper, we attempt to develop a Dirichlet Process Mixture (DPM) model to group documents into an optimal number of clusters while the number of clusters *K* is learned automatically and document clustering approach could be designed relaxing the assumption of the predefined K.

They all take the assumption that *K* is a pre-defined parameter determined by users and provided before the document clustering process. Therefore, it is useful if a document clustering approach could be designed relaxing the assumption of the pre-defined *K*. To find out the number of clusters is a difficult problem. We attempt for grouping 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 along time [1, 14, 21]. As an infinite mixture model in which each component corresponds to a different cluster, the DPM model figure out the number of clusters automatically When a new data point arrives, it either rises from existing cluster or starts a new cluster. Due to the flexibility of the DPM model it uses particularly for document clustering. However,in the some papers little work on the investigating DPM model for document clustering is done 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 non discriminative words. Only discriminative words are helpful for grouping documents.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.The first component is the discriminative words which generate from a specific topic to which document belongs. The second component is the irrelevant noise words arising from a general topic which is shared by all documents. The Dirichlet process priori is only used for the specific topics. 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 groups documents into a set of document clusters while *K* is determined automatically; and secondly 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 also contain some problems. One problem for DPM is that DPM parameters cannot be estimated quickly. To identify discriminative words, a stochastic search variable Selection technique [5, 12, 16] is applied. . In our proposed approach , the Gibbs sampling algorithm [14, 21] is used to infer both the cluster structure and the discriminative words..We compared our approach with a stage-of-the-art model based document clustering approach proposed in [9] and a standard model-based clustering approach.

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 section2. 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, we draw conclusions and make suggestions for future work.

II. 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 paper, 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 overcome on all methods which find out appropriate K to trade-off data likelihood and model complexity [11].After 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, this method is very easy to use and it aint require 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 mentioned that DPM model find out appropriate number of cluster 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] in proposed a multinomial mixture model. It implemented the EM algorithm for document clustering assuming that document topics follow multinomial distribution and each document is a combination 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 work showed that the performance of DCM was comparable to that obtained with multiple heuristic changes to the multinomial model. However, DCM model contain some problems 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 paper for document clustering in recent years.

III. 3. BACKGROUND

3.1 Dirichlet Process Mixture Model

The DPM model is nothing but a mixture model with an infinite number of mixture components [28]. 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.

Since the number of clusters is always unknown, to allow it to grow with data, we assume that the data point xn follows a general mixture model with the use of distribution G the parameter θ is generated

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. 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 Gas a (infinite-dimensional) parameter and assign a prior to it. One class of Bayesian nonparametric techniques is called the Dirichlet process (DP) [10].

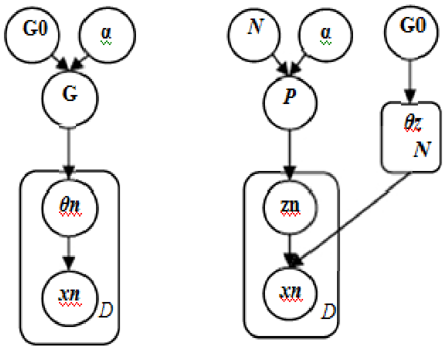

Fig 1: Graphical representation of DPM model (Left) and DMA (Right).

Dirichlet process, as a distribution on distributions, is parameterized by a positive scaling parameter α and a base distribution G0. 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:

$$\begin{aligned} &G \mid \alpha, G0 \sim DP(\alpha, G0), \\ &\theta_n \mid G \sim G, \; n=1,2,\ldots,D, \qquad (2) \\ &x_n \mid \theta_n \sim F(x_n \mid \theta_n), \; n=1,2,\ldots,D. \end{aligned}$$

The DPM model can be 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 {*θ1,…,θD*} exhibits a clustering effect. Let *θ-n* denotes the set of all *θj* for *j≠n*. 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 Cis 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. It also indicates that DP prior allows a novice data point either to share the same cluster with the previous data points or to start a new cluster. The number of clusters is figure out automatically. We can best understand this clustering property by a famous metaphor known as the Chinese restaurant process [28].

Given data points x1,. ,xD and the DP parameter (α, G0), DPM model yields a posterior distribution on θ1, …,Θd which also exhibits clustering effect [21]. 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. The clustering process which is 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:

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

where zn indicates the latent cluster allocation of the n-th sample and N is the number of mixture components. For each cluster z, the parameter θz 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 Fig1.

Let z-ndenote 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:

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

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 Ngo to infinity. It has been shown in [14] that the L1distance between the Bayesian marginal density of the data under DMA and the DPM model is O(4D exp(-(N-1)/α)). 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 L1bound 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.

IV. 4. DPMFS & 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.

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

This latent vector partitions the dataset x into two parts: first part is thediscriminative 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. We assign a prior to γand assume that its elements are independent Bernoulli random variables with common probability distribution. The distribution of γis as follows:

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

where ωis 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

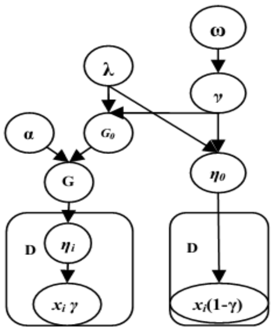

Figure 2: Graphical representation of DPMFS model.

4.2 DPM Model with Feature Selection

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

1. Choose 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. xiγ and xi (1-γ) represent (xi1γ1,…,xiWγW) and (xi1(1-γ1),…,xiW (1-γW)) respectively. ni denotes then multinomial parameter for the discriminative words in xi and η0, 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, xi γi = 1,2,…, D. Parameters in the Dirichlet distribution and Multinomial distribution used in the our model may be zero.Here we only consider those non-zero parameters. For example, the probability density functions for xi γ is as follows

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

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 xi is given by:

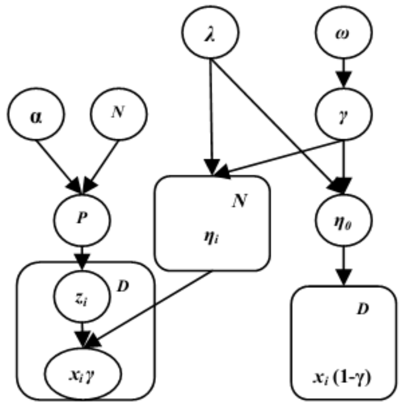

Figure 3: Graphical representation of DMAFS model.

4.3 Approximating the DPMF Model

In this section, we design a DMA model with feature selection, named DMAFS. 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 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 compare the finite mixture model. 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 η0, η1,…, ηN in Equation (10), the likelihood of the documents conditioned on the latent variables γ and z becomes:

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

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

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

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

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

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

V. ALGORITHM

Gibbs sampling method is used here to infer both the latent cluster Structure as well as discriminative words in the context of DMAFS model. The inference procedure is become more 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
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, λ and 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

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

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

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

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.This new candidate is accepted with the probability

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

5. Update λ if necessary by the following sampling:

5a. update η0by 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(λ)to λ and draw λ from

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

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 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 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 pare 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.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an approach which handles document clustering as well as feature partition simultaneously A document clustering approach is investigated based on the DPM model which groups documents into an arbitrary number of 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. Non discriminative words are regarded to be generated from a general back-ground shared by all documents. The Gibbs Sampling technique is used to infer both the cluster structure and the latent discriminative word subset. Our paper shows that DPMFS approach groups document dataset into meaningful clusters it does not require to know the number of clusters 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.

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.

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 thestandard multivariate normal mixture model using model-selection criteria.

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

[6] P. Cheeseman, J. Kelly, M. Self, JFreedman. (1988). Autoclass: A Bayesian classification system. In Proceedings of the Fifth International Conference on Machine Learning

[7] I. S. Dhillon and D. S. Modha. (2001). Concept decompositions for large sparse text data using clustering.Journal of Machine Learning,42(1):143175

[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):877-89

[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 data description.Automatica 14:465-471.

[24] K. Rose. (1998). Deterministic annealing for clustering, compression 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 cluster. 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):240-2