# Classifying Documents with Maximum Likelihood Approximation of the Dirichlet Multinomial Gibbs Model

Shibin Zhou, Zhao Cao, Yushu Liu
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, 100081, P.R. China
Email: {guoguos.zhou,zhaoyang,liuyushu}@bit.edu.cn

***Abstract*—In the text analysis, the Dirichlet compound multinomial (DCM)distribution has recently been shown to be a good model for documents because it captures the phenomenon of word burstiness, unlike the standard multinomial distribution. The burstiness phenomenon describes the behavior of a rare word appearing many times in a single document. In this paper, for the sake of improving performance of modeling documents, we propose a variant of DCM and Gibbs distribution called Dirichlet multinomial Gibbs (DMG) model by introducing Gibbs parameters to DCM distribution. We demonstrate the maximum likelihood procedure of the DMG model with these Gibbs parameters. By our experiments, the DMG approach inherit the merits of methods of Gibbs distribution approximation and DCM estimation. More specifically, as revealed by our experimental results on various real-world text datasets, we show that maximum likelihood approximation of the DMG model is more desirable than some current state-of-the-art methods.**

***Index Terms*—Document classification, Dirichlet compound multinomial model, Gibbs distribution**

## I. INTRODUCTION

Because the increased availability of documents in digital formats, there is a growing need for finding, filtering, and managing these resources. Document classification is a major solution to these kinds of problems, which is a text content-based classification technique that assigns documents to some predefined categories. In past years, a wide range of supervised learning algorithms has been applied to document classification problem, using a training set of categorized documents to build a classifier that maps arbitrary documents to relevant categories.

An important phenomenon reported in [1][2] is burstiness of words in document context. The term "burstiness" describes the behavior of words which tend to appear in bursts, i.e. once they appear in a document, they are much more likely to appear again. The notion of burstiness describes the fact that the more we find a word in a document, the higher the expectation to find new occurrences. Recently, the Dirichlet compound multinomial (DCM) [3][4] model has been shown to be well suited for modeling word burstiness in documents and count data on finite mixture distribution [5].

Dirichlet compound multinomial (DCM) is a generative approaches which are popular to classification task since they are relatively easy to interpret and can be trained quickly. With these approaches, the key problem is to develop a probabilistic model that represents the data well. In recent years there has been a proliferation of graphical models composed of a multitude of multinomial and Dirichlet variables interacting in various inventive ways.

Dirichlet Compound Multinomial model (DCM) uses a Dirichlet prior over multinomial conditionals, where Dirichlet distribution is conjugate to multinomial distribution. Like a multinomial distribution, the Dirichlet Compound Multinomial distribution is a generative model for text documents that takes into account burstiness [3] and a distribution over all possible count vectors that sum to a fixed value. Rather, the term frequency (TF) schema or multinomial distribution only figure occurrence in the document. Previous work [3] has shown that classifiers using Bayes' rule and the DCM model are competitive with the best known classification methods on standard document collections. For these Dirichlet multinomial models the inference method of choice is typically collapsed Gibbs sampling, due to its simplicity, speed, and good predictive performance on test date.

Motivating by these idea, we concern with a new approach in this paper which is a variant of DCM distribution called Dirichlet multinomial Gibbs (DMG) model by introducing Gibbs parameters to DCM distribution. In our document classification scenario, we demonstrate the maximum likelihood procedure of the DMG model. Furthermore, we show that the Gibbs parameters are similar to the parameters of features in the maximum entropy model [6]. The solution estimating on DMG model has been proposed to inherit merits of the Dirichlet compound multinomial model [3] to model word burstiness, and inherit merits of iteration approximation method of the maximum entropy text classification method [6] by our experiments. So it can model documents by not only capturing word burstiness but also approximating real texts distribution very well.


doi:10.4304/jcp.5.4.654-661

Therefore, our DMG model can improve performance of modeling documents which confirmed by our experiments conducted over various different test collections about document classification.

This paper is organized as follows. Firstly, we introduce the related concepts and works in section Ⅱ. Secondly, in section Ⅲ we describe Dirichlet Multinomial Gibbs (DMG) model in detail. Next, in section Ⅳ the method estimating parameters of DMG model will be derived. While section Ⅴ we show the experimental results on various datasets using DMG model to document classification. Finally, we conclude the paper with a summary in section Ⅵ.

## II. Related concepts and works

### A. Burstiness property

The term "burstiness" [1][2] describes the behavior of a rare word appearing many times in a single document. Due to the large number of possible words, most words do not appear in a given document. Nevertheless, if one word does appear once, it is much more likely to appear again, i.e. words appear in bursts.

Let $\mathrm{tf}_w$ be the number of occurrences of word $w$ in a given document. For a word probability distribution $p$, [2] measures the burstiness through the quantity

$$B_p = \frac{\mathrm{E}_p\left[\mathrm{tf}_w\right]}{p\left(\mathrm{tf}_w \geq 1\right)}$$

with $B_p$ denotes the expectation with respect to distribution $p$. In order to give a clear measure on whether a given word distribution accounts or not for bursty and non-bursty words. [7] introduced a different definition. Say that a word $w$ is bursty at level $n_0$ $1 \leq n_0$, such that for all integers $(n', n)$, $n' \geq n \geq n_0$:

$$p\left(\mathrm{tf}_w \geq n'+1 \mid \mathrm{tf}_w \geq n'\right) \geq p\left(\mathrm{tf}_w \geq n+1 \mid \mathrm{tf}_w \geq n\right)$$

This definition directly translate the fact that a word is bursty if it is easier to generate it again once it has been generated a certain number of times (passed a certain level). The introduction of a burstiness level ($n_0$) in this definition allows one to capture finer-grain behavior.

The multinomial captures the burstiness of common words, but the burstiness of average and rare words is not modeled correctly. This is a major deficiency in the multinomial model since rare and average words represent 99% of the vocabulary and 29% of emissions [3] and, more importantly, these words are key features for classification. An explanation for this behavior is that the common words are more likely to satisfy the independence assumption, since many of the common words are non-content, function words. The rare and average words are information-carrying words, making them more likely to appear if they have already appeared in a document.

### B. The Dirichlet Compound Multinomial Distribution

Dirichlet Compound Multinomial model (DCM) [3] uses a Dirichlet prior over multinomial conditionals, where Dirichlet distribution is conjugate to mutinomial distribution. Like a multinomial distribution, the Dirichlet Compound Multinomial distribution is a generative model for text documents that takes into account burstiness [3] and a distribution over all possible count vectors that sum to a fixed value. Rather, the term frequency (TF) schema or multinomial distribution only figure occurrence in the document.

Suppose $d$ is a document, The DCM distribution [3] is

$$p(d \mid \alpha) = \frac{\left(\sum_{t=1}^{n} \mathrm{tf}_{t,d}\right)!}{\prod_{t=1}^{n} \mathrm{tf}_{t,d}!} \frac{\Gamma\left(\sum_{t=1}^{n} \alpha_t\right)}{\Gamma\left(\sum_{t=1}^{n} \mathrm{tf}_{t,d} + \alpha_t\right)} \times \prod_{t=1}^{n} \frac{\Gamma\left(\mathrm{tf}_{t,d} + \alpha_t\right)}{\Gamma\left(\alpha_t\right)}$$

Where $\mathrm{tf}_{t,d}$ stores the number of occurrences of a word $t$ in document $d$ and $\Gamma$ is gamma function.

So given a corpus $D$ of documents, the dirichlet prior parameter $\alpha$ can be derived by following iteration [8]:

$$\alpha_k^{\mathrm{new}} = \alpha_k \frac{\sum_{d \in D} \psi\left(\mathrm{tf}_{t,d} + \alpha_k\right) - \Gamma\left(\alpha_k\right)}{\sum_{d \in D} \psi\left(\sum_{t=1}^{n} \mathrm{tf}_{t,d} + \sum_{t=1}^{n} \alpha_t\right) - \psi\left(\sum_{t=1}^{n} \alpha_t\right)}$$

where $\psi$ is digamma function.

Like a multinomial distribution, a DCM is a distribution over all possible count vectors that sum to a value $\sum_{t=1}^{n} \mathrm{tf}_{t,d}$. When a DCM or a multinomial is used to model a collection of documents of different lengths, formally there is a different distribution for each different length, with all distributions sharing the same parameter values. Also, with a DCM or with a multinomial, $p(d)$ for a document $d$ is really the probability of the equivalence class of all documents that have the same word counts, that is all documents that have the same bag-of-words representation.

Moreover, the DCM model is a generative model for the documents within a class and it can represent a topic where different documents use alternative terminology [3]. This within-topic diversity is different from the within-document diversity allowed by latent topic modeling, where each topic is represented by a single multinomial, but each word in a document may be generated by a different topic.

### C. Gibbs distribution

Gibbs distribution [9] related heavily to Boltzmann distribution. Boltzmann distribution originated from Liouville theorem. It is just one step to the Gibbs distribution, the corner stone of the equilibrium Statistical Mechanics. Imagine a large mechanical system with virtually infinite number of microscopic degrees of freedom. This system is supposed to be in equilibrium in the sense that all its macroscopic motions have relaxed, so that the time evolution is nothing but repeating

microscopically different, but macroscopically equivalent states. We will refer to such a system as a heat bath.

Boltzmann distribution consider systems which are in contact with a heat bath at temperature $T$ and also in with a particle reservoir at chemical potential $\mu$. The temperature is a measure of the decrease in entropy of the reservoir from giving up heat to the system (see here); the chemical potential is a measure of the energy decrease (and entropy increase) of the reservoir from giving up particles to the system. Boltzmann distribution want to find the probability that our system will be in a certain microstate $i$ with an energy $\varepsilon_i$ and particle number $N_i$.

The derivation follows that of the Boltzmann distribution closely. Again the probability of the system being in the given microstate depends on the number of microstates available to the reservoir with energy $E_0 - \varepsilon_i$ and particle number $N_0 - N_i$. Expressing the number of microstates as the exponential of the entropy, making a Taylor expansion of the entropy about $S_R(E_0 - N_0)$, and expressing the derivatives of the entropy in terms of $T$ and $\mu$ thus,

$$\left(\frac{\partial S_R}{\partial E}\right)_{V,N} = -\frac{1}{T}, \quad \left(\frac{\partial S_R}{\partial E}\right)_{E,N} = -\frac{\mu}{T}$$

gives

$$p_i = \frac{1}{Z}\exp\left(\frac{\mu N_i - \varepsilon_i}{k_B T}\right)$$

with

$$Z = \sum_i \exp\left(\frac{\mu N_i - \varepsilon_i}{k_B T}\right)$$

where the parameter $k_B$ is called Boltzmann factor. The normalization constant $Z$ is called the partition function. Macroscopic functions of state are calculated via ensemble averages as usual; the relevant ensemble in this case is called the canonical ensemble.

## III. DMG MODEL

### A. Notation

We describe some notations using in Dirichlet multinomial Gibbs (DMG) model. Suppose we have $N$ documents which can be classified to $M$ categories, containing words from corpus $D$ who has a vocabulary $V$ of size $v$. The corpus of text documents is summarized in a $N$ by $v$ co-occurrence table, where $x_{t,d}$ stores the number of occurrences of a word $t$ in document $d$. We would like to use $p(d \mid \theta_c)$ to denote probability of generating document $d$ from a multinomial distribution with parameters vector $\theta_c$ specified to category $c$, $p(\theta_c \mid \alpha)$ to denote the probability of vector $\theta_c$ with Dirichlet distribution parameters vector $\alpha$.

### B. Multinomial modeling of text

When using a multinomial distribution to model text, the multinomialdistribution specifies the probability of observing a given vector of word counts, where the probabilit $\theta_{t,c}$ of emitting word $t$ is subject to the constraints $\sum_{t\in V}\theta_{t,c} = 1$ and $\theta_{t,c}$ for all $t$ specified to category $c$. The probability of a document $d$ generated by this model is

$$p(d \mid \theta_c) = \frac{n!}{\prod_{t\in V} x_{t,d}!}\prod_{t\in V}\theta_{t,c}^{x_{t,d}} \tag{1}$$

where $n = \sum_{t\in V} x_{t,d}$.

### C. Dirichlet modeling of text

The majority of the researchers assign a Dirichlet prior to the parameter vector of a multinomial distribution. This is due to the fact that the Dirichlet is a conjugate prior to the multinomial distribution, that is, the posterior is also Dirichlet. It is defined as

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{t\in V}\alpha_t\right)}{\prod_{t\in V}\Gamma(\alpha_t)}\prod_{t\in V}\theta_t^{\alpha_t - 1} \tag{2}$$

### D. DCM modeling of text

Given documents set $D_c$ of category $c$, we get likelyhood function which is also Dirichlet distribution as

$$p(D_c, \theta_c \mid \alpha) = \frac{\Gamma\left(\sum_{t\in V} x_{t,D_c} + \alpha_t\right)}{\prod_{t\in V}\Gamma\left(x_{t,D_c} + \alpha_t\right)}\prod_{t\in V}\theta_{t,c}^{x_{t,D_c} + \alpha_t - 1} \tag{3}$$

We know expectation of $\theta_c$ according to [10] about this Dirichlet distribution is

$$\mathrm{E}(\theta_{t,c}) = \frac{x_{t,D_c} + \alpha_t}{\sum_{t\in V} x_{t,D_c} + \alpha_t} \tag{4}$$

Thus, the maximum likelihood of $\theta_c$ is

$$\theta_{t,c} = \frac{x_{t,D_c} + \alpha_t}{\sum_{t\in V} x_{t,D_c} + \alpha_t} \tag{5}$$

So, given a new document $d$, we can estimate probability of document $d$ with respect to $\theta_c$ according Eq.(1). Eventually, by Dirichlet compound multinomial (DCM) modeling, we can get posterior probability of the document $d$ belong to category $c$ as

$$p_{\mathrm{dcm}}(c \mid d) \propto p(d \mid \theta_c)\, p(c) \tag{6}$$

where $p(c)$ is probability of category $c$. So, according to Eq.(6) and (1), we have

$$p_{\mathrm{dcm}}(c \mid d) \propto p(c)\prod_{t\in V}\theta_{t,c}^{x_{t,d}} \tag{7}$$

### E. DMG modeling of text

We can represent Eq.(1) by formulation of Gibbs distribution [11] as

$$p_{\mathrm{dcm}}(c \mid d) \propto p(c)\exp\left(\sum_{t\in V} x_{t,d}\log\theta_{t,c}\right) \tag{8}$$

Establishing connections between Gibbs and Dirichlet compound multinomial, we think that it will allow to combine virtues of Gibbs distribution and DCM model. Along this line, we introduce Gibbs parameters $\lambda \in \Lambda$ into DCM distribution and composite the Dirichlet multinomial Gibbs (DMG) model. So we represent DMG distribution as

$$p_\Lambda(c \mid d) = \frac{1}{Z_\Lambda(d)} p(c) \exp\left( \sum_{t \in V} \lambda_{t,c} x_{t,d} \log \theta_{t,c} \right) \quad (9)$$

where $Z_\Lambda(d) = \sum_c p(c) \exp\left( \sum_{t \in V} \lambda_{t,c} x_{t,d} \log \theta_{t,c} \right)$.

Hereafter, we have $p(c \mid d)$ as maximum likelihood of $p_\lambda(c \mid d)$ with respect to Gibbs parameters $\lambda$.

$$p(c \mid d) = \max_\lambda p_\lambda(c \mid d) \quad (10)$$

*F. The analysis of DMG model*

According to above description, DMG model make up of random field [12]. In the following, we analyse the random field property of DMG model according to [12].

The random field model can be manifest by a finite graph $G$ [12]. Let $G = (E, V)$ with vertex set $V$ and edge set $E$. Configuration space $\Omega$ can be defined as the set of the vertices in $V$ which have the same label. For DMG model, $c$ is a configuration. If a clique $cl \subset V$ and $\omega \in \Omega$ is a configuration, then $\omega_{cl}$ denotes the configuration restricted to $cl$. A random field on $G$ is a probability distribution on $\Omega$. We can define configuretion function $f : \Omega \to R$ and we assume that each $cl$ is a path-connected subset of $V$ and that

$$V_{cl}(\omega) = \sum_{1 \le i \le n_{cl}} \lambda_i^{cl} f_i^{cl}(\omega) = \lambda^{cl} \cdot f^{cl}(\omega)$$

where $\lambda_i^{cl} \in R$ and the values $\lambda_i^{cl}$ are the parameters of the DMG field, the functions $f_i^{cl}$ are the features of the field. It will usually regard these features and parameters independent on a vertex clique $cl$. And the field can be express in the form

$$p(\omega) = \frac{1}{Z} \exp\left( \sum_i \lambda_i f_i(\omega) \right) = \frac{1}{Z} \exp\left( \lambda \cdot f(\omega) \right)$$

where Z is normalization in order to make $p(\omega)$ is a probability function.

Hence it will be convenient to express our DMG model in terms of the random field models described above according Eq.(9). There are several contact points. The Gibbs distribution can be derived by maximizing entropy: basically, it has maximal entropy among all probability measures on $\Omega$ with the same average. We should also like to mention the interesting observation to that conventional ME restoration is a special case of MAP estimation in which the prior distribution on F [9] is

$$p(\omega) = \frac{1}{Z} \exp\left( -\beta \sum_{i,j} \omega_{i,j} \log \omega_{i,j} \right)$$

which is a variant of Eq.(9). By conventional maximum entropy method, we prefer to maximizing the entropy $f_{i,j} \log f_{i,j}$.

In the following, we discuss two optimization problems [12] of the random fields which related to our DMG model. Suppose that we are a set of features $f = (f_1, f_2, \cdots, f_n)$, and $\tilde{p}$ is the empirical distribution of a set of training documents $d_1, d_2, \cdots, d_N$. We can derive a probability distribution $q^*$ that accounts for these training documents and approximates $\tilde{p}$.

Usually, we measure distance between probability distributions $p$ and $q$ using the Kullback-Leibler divergence

$$D(p \,\|\, q) = \sum p \log \frac{p}{q}$$

and we have the expectation of function or feature $g$ about distribution $p$ as

$$\mathrm{E}_p[g] = \sum g \cdot p$$

There are two natural sets of probability distributions determined by the data $\tilde{p}$ and $f$. The first is the set $P(f, \tilde{p})$ of all distributions that agree with $\tilde{p}$ as to the expected value of the feature function $f$:

$$P(f, \tilde{p}) = \left\{ p : \mathrm{E}_p[f] = \mathrm{E}_{\tilde{p}}[f] \right\}$$

The second is the set $Q(f)$ of generalized Gibbs distributions based on feature function $f$:

$$Q(f) = \left\{ \exp(\lambda f) : \lambda \in R^n \right\}$$

We let $\bar{Q}(f)$ denote the closure of $Q(f)$.

There are two natural criteria for choosing an element $q^*$ from these sets: one is maximum likelihood method and the other is maximum entropy method.

Maximum Likelihood Gibbs Distribution. Choose $q^*$ to be a distribution in $\bar{Q}(f)$ with maximum likelihood with respect to $\tilde{p}$:

$$q_*^{\mathrm{ML}} = \arg\min_{q \in \bar{Q}(f)} \sum \tilde{p} \log \frac{\tilde{p}}{q}$$

Maximum Entropy Constrained Distribution. Choose $q^*$ to be a distribution in $P(f, \tilde{p})$ that has maximum entropy:

$$q_*^{\mathrm{ME}} = \arg\min_{p \in P(f, \tilde{p})} \sum p \log p$$

These criteria are different, but they determine the same distribution: $q^* = q_*^{\mathrm{ML}} = q_*^{\mathrm{ME}}$. Furthermore, this distribution is the unique element of the intersection of sets $P(f, \tilde{p}) \cap \bar{Q}(f)$. In conclusion, for DMG the maximum likelihood of the model equal to maximum entropy of the model.

## Ⅳ. ESTIMATING PARAMETERS OF DMG

In section Ⅲ, we get representation of DMG model and we also know it is a variant of DCM and Gibbs distribution. Therefore, we can obtain maximum likelihood approximation of DMG model by using improved iterative scaling (IIS)[12] algorithm to estimate Gibbs parameter. In this section, we briefly outline the derivation of IIS which is a hillclimbing algorithm. A complete description and derivation of improved iterative scaling method is presented by [12]. We describe the algorithmic details of this procedure follows [13] and [6]. We know that IIS performs hillclimbing in parameter log likelihood space. Then, given a set of i.i.d. training data $D$, we can designate the log likelihood $L(\Lambda \mid D)$ of an DMG model with regard to $\Lambda$ which is the set of the gibbs parameters:

$$L(\Lambda \mid D) = \log \prod_{d \in D} p_\Lambda (c \mid d)$$
$$= \sum_{d \in D} \log p(c) \sum_{t \in V} \lambda_{t,c} x_{t,d} \log \theta_{t,c} \quad (11)$$
$$- \sum_{d \in D} \log \left[ \sum_c p(c) \exp \left( \sum_{t \in V} \lambda_{t,c} x_{t,d} \log \theta_{t,c} \right) \right]$$

where $\lambda_{t,c} \in \Lambda$.

Let us define

$$f_{t,c'}(d,c) = \begin{cases} x_{t,d} \log \theta_{t,c} & c' = c \\ 0 & c' \neq c \end{cases} \quad (12)$$

Then at each step, IIS algorithm can find an incrementally more likely set of parameters which denote by $\Delta$. When starting from some initial vector parameters $\Lambda$, we improve $\Lambda$ by setting it equal to $\Lambda + \Delta$. Thus we expect to obtain a $\Delta$ such that the difference in likelihoods is positive:

$$L(\Lambda + \Delta \mid D) - L(\Lambda \mid \Delta) \geq 0 \quad (13)$$

Referring to Eq.(11) we have

$$L(\Lambda + \Delta \mid D) - L(\Lambda \mid \Delta)$$
$$= \sum_{d \in D} \left( \log p(c) \sum_i \delta_i f_i(d,c) \right) \quad (14)$$
$$- \sum_{d \in D} \log \left[ \sum_c \frac{p(c) \exp\left( \sum_i (\lambda_i + \delta_i) f_i(d,c) \right)}{\sum_c p(c) \exp\left( \sum_i \lambda_i f_i(d,c) \right)} \right]$$

where $\lambda_i \in \Lambda$ and $\delta_i \in \Delta$.

And according to Eq.(9), above equation (14) also can be represented as

$$L(\Lambda + \Delta \mid D) - L(\Lambda \mid \Delta)$$
$$= \sum_{d \in D} \left( \log p(c) \sum_i \delta_i f_i(d,c) \right) \quad (15)$$
$$- \sum_{d \in D} \log \left[ \sum_c p_\Lambda (c \mid d) \exp\left( \sum_i \delta_i f_i(d,c) \right) \right]$$

Using the inequality $\log x \leq x - 1$, we can get:

$$L(\Lambda + \Delta \mid D) - L(\Lambda \mid \Delta)$$
$$\geq \sum_{d \in D} \left( \log p(c) \sum_i \delta_i f_i(d,c) \right) + \sum_{d \in D} 1 \quad (16)$$
$$- \sum_{d \in D} \left[ \sum_c p_\Lambda (c \mid d) \exp\left( \sum_i \delta_i f_i(d,c) \right) \right]$$

We can apply Jensen's inequality -- namely for a p.d.f. $p(x)$

$$\exp \sum_x p(x) q(x) \leq \sum_x p(x) \exp q(x)$$

So we can bound this expression as:

$$L(\Lambda + \Delta \mid D) - L(\Lambda \mid \Delta)$$
$$\geq \sum_{d \in D} \left( \log p(c) \sum_i \delta_i f_i(d,c) \right) + \sum_{d \in D} 1 \quad (17)$$
$$- \sum_{d \in D} \left[ \sum_c p_\Lambda (c \mid d) \sum_i \frac{f_i(d,c)}{f_\#(d,c)} \exp\left( \delta_i f_\#(d,c) \right) \right]$$

where $f_\#(d,c) = \sum_i f_i(d,c)$.

Let us define an auxiliary function called $B$ to denote right hand side of above equation (17):

$$B = \sum_{d \in D} \left( \log p(c) \sum_i \delta_i f_i(d,c) \right) + \sum_{d \in D} 1$$
$$- \sum_{d \in D} \left[ \sum_c p_\Lambda (c \mid d) \sum_i \frac{f_i(d,c)}{f_\#(d,c)} \exp\left( \delta_i f_\#(d,c) \right) \right] \quad (18)$$

If $B \geq 0$, we can gaurantee an increase in the likelihood. We can derive optimal $\Delta$ by differentiating $B$ with respect to the trivial changes in each parameter $\delta_i \in \Delta$ and solving for the maxima:

$$\frac{\partial B}{\partial \delta_k} = \sum_{d \in D} \log p(c) \delta_k f_k(d,c)$$
$$- \sum_{d \in D} \left[ \sum_c p_\Lambda (c \mid d) f_k(d,c) \exp\left( \delta_k f_\#(d,c) \right) \right] \quad (19)$$

This analysis demonstrate that the likelihood is convex. Thus, we can improve the model likelihood at each hillclimbing step, such as Newton-Raphson method, which is guarenteed to converge to the global maximum.

So the procedure of IIS algorithm for estimating parameters of DMG model can be shown as follows:

- Given A collection $D$ of labeled documents and a set of feature functions $f_i$.
- Our target is estimating every feature $f_i$ expected value on the training documents by the constraints Eq.(12).
- Initialize all the parameters $\lambda_i$ 's to be zero.
- Iterate the model likelihood until convergence by Newton-Raphson method:
  - Calculate the expected class labels for each document with the current parameters according to Eq.(9).
  - For each parameter $\lambda_i$ :
    - Set $\partial B / \partial \delta_i$ and solve for $\delta_i$ according to Eq.(19).
    - set $\lambda_i = \lambda_i + \delta_i$.
- At last, we can derived a document classifier that takes an unlabeled document and predicts a class label.

## Ⅴ. Experiments and Results

In order to evaluate the properties of our method, we have conducted experiments on two real-world datasets, WebKB and 20Newsgroups, to evaluate the effectiveness of our proposed model for text categorization.

### A. Datasets

The 20Newsgroups(20NG)[1] dataset is a collection of approximately 20,000 documents that were collected from 20 different newsgroups with about 1000 messages from each newsgroup. This collection consists of 19,974 non-empty documents distributed evenly across 20 newsgroups and we selected 19,946 nonempty documents which are all the same after feature selection. We use the newsgroups to form categories, and randomly select 70% of the documents to be used for training and the remaining 30% for testing.

The WebKB[2] dataset contains manually classified Web pages that were collected from the computer science departments of four university (Cornell, Texas, Washington and Wisconsin) and some other university. The pages are divided into seven categories: student, faculty, staff, course, project, department and other. In this paper, we use the four most populous entity-representing categories: student, faculty, course, and project, which all together contain 4199 pages. We called this selected WebKB dataset as WebKB top-4 dataset. Like handling 20Newsgroups dataset, We randomly select 70% of the documents to be used for training and the remaining 30% for testing.

### B. Gaussian Prior of DMG Model

Dirichlet multinomial Gibbs (DMG) model can also suffer from overfitting like document classification method proposed by Nigam et al [6]. Chen and Rosenfeld [14] have shown applying a Gaussian prior centered around to smooth the maximum likelihood, which can ameliorate over-fitting. Goodman [15] have done the similar work also. By performing maximum a posteriori instead of maximum likelihood estimation toward the uniform model, the Gaussian prior method adds little computation to existing maximum likelihood estimation algorithms. Since the logarithm of the Gaussian prior is concave the objective function is still concave, we can make a simple modification to improved iterative scaling (IIS) to find the MAP model. Accordingly, we integrate a Gauss prior into DMG model with the mean at zero and a diagonal covariance matrix.

The prior probability of the model is just the product over the Gaussian of each feature value $\lambda_i$ with variance $\sigma_i^2$ :

$$p(\Lambda) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\lambda_i^2}{2\sigma_i^2}\right)$$

Integrating this prior into improved iterative scaling requires adding a single term to the derivative of $B$ (Equation 19):

$$\frac{\partial B}{\partial \delta_k} = \frac{\lambda_i + \delta_i}{-\sigma_i^2} + \sum_{d \in D} \log p(c) \delta_k f_k(d,c) - \sum_{d \in D} \left[ \sum_c p_\Lambda(c \mid d) f_k(d,c) \exp\left(\delta_k f_\#(d,c)\right) \right] \quad (20)$$

[1] http://people.csail.mit.edu/jrennie/20newsgroups

2 http://people.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data

TABLE I.
EXPERIMENTAL RESULTS ON THE WEBKB TOP-4

| | SVM | MaxEnt | DMG |
|---|---|---|---|
| macro-averaging precision | 0.753 | 0.892 | 0.889 |
| macro-averaging recall | 0.703 | 0.879 | 0.865 |
| macro-averaging F1 | 0.720 | 0.855 | 0.876 |
| micro-averaging accuracy | 0.733 | 0.897 | 0.891 |

TABLE II.
EXPERIMENTAL RESULTS ON THE 20NG DATASET

| | SVM | MaxEnt | DMG |
|---|---|---|---|
| macro-averaging precision | 0.817 | 0.803 | 0.824 |
| macro-averaging recall | 0.801 | 0.804 | 0.823 |
| macro-averaging F1 | 0.805 | 0.803 | 0.823 |
| micro-averaging accuracy | 0.798 | 0.801 | 0.823 |

Moreover, this new formula is easily solved for a maximum with a numeric root-finding procedure, like Newton's method. [14] have shown that introducing a Gaussian prior on each $\lambda_i$ improves performance for language modeling tasks when sparse data causes over-fitting. This paper also derives the update rule give by above equation.

### C. Experiments

The standard performance measure for document classification systems is precision and recall. Precision measures how many of the retrieved entries are relevant. Recall measures how many relevant entries were found compared to the amount of relevant entries in the collection. The $F_1$ measure [16] weights the importance of precision and recall equally.

TABLE III.
THE COMPARISON OF PERFORMANCE (F1) ON WEBKB TOP-4

| Category | SVM | MaxEnt | DMG |
|---|---|---|---|
| student | 0.878 | 0.917 | 0.912 |
| faculty | 0.817 | 0.870 | 0.857 |
| course | 0.446 | 0.931 | 0.947 |
| project | 0.741 | 0.818 | 0.782 |
| macroaveraging F1 | 0.720 | 0.885 | 0.876 |

Macro averaging [16] evaluate precision and recall for each category, and average over the results of the different categories. However, micro averaging [16] over all the classes, is rewarded when classifiers of frequent categories performs well. So the micro averaged precision and recall simplifies to the fraction of correct classified documents. It follows from that the $F_1$ measure also becomes the fraction of correct classified documents. These two methods may give quite different results, especially if the different categories have very different

TABLE IV.
THE COMPARISON OF PERFORMANCE (F1) ON 20NG SUBSET

| Category | SVM | MaxEnt | DMG |
|---|---|---|---|
| alt.atheism | 0.708 | 0.697 | 0.723 |
| comp.graphics | 0.736 | 0.759 | 0.786 |
| comp.os.ms-windows.misc | 0.806 | 0.791 | 0.792 |
| comp.sys.ibm.pc.hardware | 0.753 | 0.723 | 0.730 |
| comp.sys.mac.hardware | 0.815 | 0.778 | 0.822 |
| comp.windows.x | 0.830 | 0.812 | 0.851 |
| misc.forsale | 0.607 | 0.767 | 0.791 |
| rec.autos | 0.877 | 0.854 | 0.878 |
| rec.motorcycles | 0.895 | 0.909 | 0.917 |
| rec.sport.baseball | 0.934 | 0.936 | 0.960 |
| rec.sport.hockey | 0.948 | 0.943 | 0.954 |
| sci.crypt | 0.884 | 0.894 | 0.926 |
| sci.electronics | 0.766 | 0.771 | 0.797 |
| sci.med | 0.897 | 0.874 | 0.899 |
| sci.space | 0.926 | 0.913 | 0.920 |
| soc.religion.Christian | 0.875 | 0.877 | 0.873 |
| talk.politics.guns | 0.803 | 0.760 | 0.808 |
| talk.politics.mideast | 0.893 | 0.879 | 0.898 |
| talk.politics.misc | 0.659 | 0.623 | 0.657 |
| talk.religion.misc | 0.480 | 0.480 | 0.496 |
| macroaveragingF1 | 0.805 | 0.803 | 0.823 |

generality. For instance, the ability of a classifier to behave well also on categories with low generality (i.e., categories with few positive training instances) will be emphasized by macro averaging and much less so by micro averaging. So, in this experiment, we just focus on macro precision, macro recall, macro $F_1$ and micro accuracy (precision) measure.

For these two datasets, we performed stop word removal, stemming, and case-conversion to lower case before feature selection was applied on the training set. Furthermore, We apply Information Gain feature selecting method to the documents of both WebKB and 20NG datasets with threshold -0.0436 to WebKB and 0.055 to 20NG.

We deployed LIBSVM [17] implementation of SVM which uses the "one vs rest" method for multi-category classification because of its effectiveness and efficiency. We set all the parameters to their default values except few parameters were set special values for various datasets. Especially, we adapted LIBSVM with polynomial kernel to execute document classification tasks.

We implement standard maximum entropy according to the work of Nigam et al [6]. And we use free software MALLET [18] for the maximum entropy method of document classification tasks. Developed by Andrew McCallum, MALLET is a library of Java code for

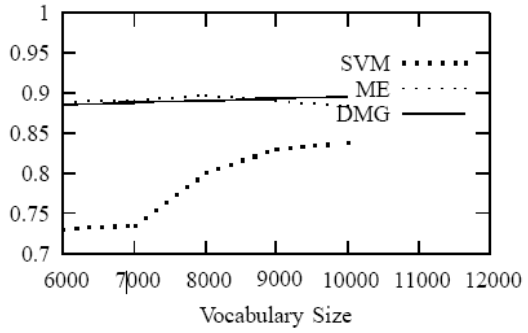


Figure 1. The micro-averaging precision on the WebKB data set with different vocabulary size according to SVM, MaxEnt and DMG model.

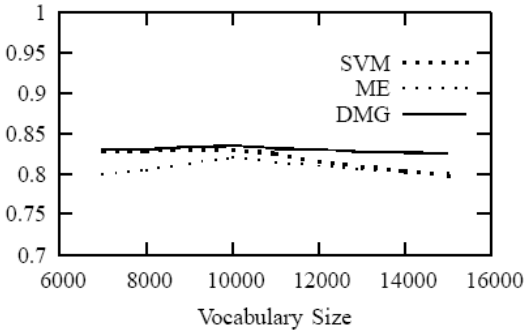


Figure 2. The micro-averaging precision on the 20newsgroups data set with different vocabulary size according to SVM, MaxEnt and DMG model.

machine learning applied to text. It provides facilities for many natural language processing, such as document classification.

The results of macro-averaging and micro-averaging to WebKB and 20NG datasets are shown in Tables Ⅰ and Ⅱ for standard maximum entropy (MaxEnt), SVM with polynomial kernel(SVM) and Dirichlet multinomial Gibbs (DMG) model respectively. We also showed the comparison of performance $F_1$ on WebKB and 20NG subsets in Table Ⅲ and Table Ⅳ. The micro-averaging accuracy comparison between our algorithm and other methods according to different vocabulary size are shown in Fig. 1 and Fig. 2 for WebKB and 20newsgroups respectively. Specially, All results are averaged across 5 random runs for WebKB and 20NG dataset.

According experimental results, Dirichlet multinomial Gibbs (DMG) model has better effects on both datasets. And we can affirm that our model has potential energy to model document in area of text analysis. In the future work, we figure to modify the method of iterating parameters in order to obtain more perfect approximation.

## Ⅵ. CONCLUSION

In this paper we derived a new approach modeling documents called Dirichlet multinomial Gibbs (DMG) model, which inherit merits of the Dirichlet compound multinomial (DCM) modeling word burstiness. We have shown that the maximum likelihood estimation method of the DMG model with Gauss prior which smooth the likelihood to overcome overfitting problem, and the

method approximate real texts distribution very well. We have demonstrated experiements using our DMG model on WebKB and 20Newsgroups datasets. According to the experimental results, Dirichlet multinomial Gibbs(DMG) model which we proposed is more desirable than SVM with polynomial kernel and standard maximum entropy.

## ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their valuable comments and suggestions. We are grateful for Shidong Feng's helpful discussion and advice. Many thanks also give to Jian Cao, Jinghua Bai, Xu Zhang and Yingfan Gao for their suggestions regarding this paper. This work was supported by the Pre-Research Project of the `Eleventh Five-Year-Plan' of China under grant No.200504123.

## REFERENCES

[1] K. Church and W. Gale, "Poisson mixtures," *Natural Language Engineering*, vol. 1, no. 2, p. 163C190, 1995.
[2] S. Katz, "Distribution of content words and phrases in text and language modeling," *Natural Language Engineering*, vol. 2, no. 1, pp. 15–59, 1996.
[3] R. E. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the Dirichlet distribution," in *Proceedings of the* 22nd *International Conference on Machine Learning*, New York, NY, USA, 2005, pp. 545–552.
[4] C. Elkan, "Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution," in *Proceedings of the* 22rd *International Conference on Machine Learning*, Pittsburgh, PA, USA, 2006, pp. 289–296.
[5] N. Bouguila, "Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 462–474, 2008.
[6] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholm, Sweden, 1999, pp. 61–67.
[7] S. Clinchant and E. Gaussier, "The BNB Distribution for Text Modeling," in *Proceedings of the 30th annual European Conference on Information Retrieval Research*, Glasgow, UK, 2008, p. 150C161.
[8] T. Minka, "Estimating a dirichlet distribution," 2003, unpublished paper available at http://research.microsoft.com/˜minka.
[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–742, 1984.
[10] J. Aitchison, *The statistical analysis of compositional data*. London, UK: Chapman and Hall, 1986.
[11] H. Derin and P. Kelly, "Discrete-Index Markov-Type Random Processes," *Proceedings of IEEE*, vol. 77, no. 10, pp. 1485–1510, 1989.
[12] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380-393, 1997.
[13] A. Berger, "Convexity, maximum likelihood and all that," 1998, unpublished paper http://www.cs.cmu.edu/˜aberger.
[14] S. Chen and R. Rosenfeld, "A Gaussian Prior for Smoothing Maximum Entropy Models," School of Computer Science, Carnegie Mellon University, Technical Report CMUCS-99-108, 1999.
[15] J. Goodman, "Exponential priors for maximum entropy models," 2008, US Patent 7,340,376.
[16] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, p. 1-47, 2002.
[17] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001, software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.
[18] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002, http://mallet.cs.umass.edu.

**Shibin Zhou** received the BS degree in automation engineering from the Daqing Petroleum Institute in 1994. He received the MS degree in computer science and engineering from Beijing Institute and Technology in 2005. He is currently PhD candidate of Beijing Institute and Technology majoring in computer science and Technology. His research interests are text classification, clustering and machine learning .

**Zhao Cao** received the BS degree in computer science and engineering from Beijing Institute and Technology in 2004. He is currently PhD candidate of Beijing Institute and Technology majoring in computer science and Technology. And he is also a visiting scholar at University of Massachusetts now. His current research interests are database schema match and flash memory storage technology.

**Yushu Liu** is a supervisor of PhD candidate in Beijing Institute of Technology. Since 1991, he has been a professor of computer science and engineering at Beijing Institute of Technology. Over the last 30 years, he has published numerous book chapters and peer-reviewed journal and conference papers. His current research interests are artificial intelligence, machine learning, information processing.