Correlated Tag Learning in Topic Model

Shuangyin Li[*], **Rong Pan**[†], **Yu Zhang**[*] **and Qiang Yang**[*]
[*]Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China
{shuangyinli, zhangyu, qyang}@cse.ust.hk
[†]School of Data and Computer Science, Sun Yat-sen University, Guang Zhou, China. panr@sysu.edu.cn

Abstract

It is natural to expect that the documents in a corpus will be correlated, and these correlations are reflected by not only the words but also the observed tags in each document. Most previous works model this type of corpus, which are called the semi-structured corpus, without considering the correlations among the tags. In this work, we develop a Correlated Tag Learning (CTL) model for semi-structured corpora based on the topic model to enable the construction of the correlation graph among tags via a logistic normal participation process. For the inference of the CTL model, we devise a variational inference algorithm to approximate the posterior. In experiments, we visualize the tag correlation graph generated by the CTL model on the DBLP corpus and for the tasks of document retrieval and classification, the correlation graph among tags is helpful to improve the generalization performance compared with the state-of-the-art baselines.

1 INTRODUCTION

Documents are usually composed of a group of words with different word frequencies, leading to the 'bag-of-words' representation. Besides, it is natural to expect that documents in a corpus are highly correlated with each other. This implicit relationship among documents may be embodied in the semantic meanings of the words in each document, where we can use topic models or other related methods to learn the correlations. However, most of the documents contain not only unstructured contexts (e.g., the plain text) but also metadata (e.g., tags). The metadata usually consists of several tags, such as authors in an article, keywords for a web page, and categories for a product. To model this type of documents which are called semi-structured documents, the metadata information would play an important role in organizing, understanding, and summarizing them in many applications.

Obviously, the tags in a corpus come from a compacted space, taking higher-level semantic as one type of semantic abstraction than words. Thus, the tags should be highly correlated with each other, which is consistent with documents' correlations. That is, the correlation between two documents can be reflected via their tags. Thus, modeling the correlations among tags can benefit the learning of the relations among documents and help obtain more meaningful representations for documents, which can be helpful for the consequent tasks such as document classification and retrieval. On the other hand, to model the documents, only considering the word information is obviously not enough if the tag information is available. Meanwhile, ignoring the correlations among tags is deficient, because the correlations can help understand the documents in a better way. Hence, how to model the correlations among tags together with the words is interesting and important for document modeling.

In fact, tags can be treated as high-level 'topics' in a corpus. While differently, the observed tags would be very complicated and high-dimensional, and belong to a different semantic space, compared with latent topics discovered by topic models. Thus, there should be a connection between the observed tags and the latent topics, such as a distribution over topics for each tag. In previous works such as the tag-weighted topic model [15], the author topic model [18], and the labeled-LDA [21], almost all of them define continuous distributions for the observed tags over the latent topics. Each tag is defined as a vector sampled from a certain probability distribution such as a Dirichlet distribution in [15], where the vector for a tag indicates the distribution over all the latent topics. In this way, the observed tags and the latent topics are combined together. However, under the Dirichlet distribution, the tags are modeled to be independent, which ignores the correlations among the tags.

On the other hand, we may model the correlations among the tags only using the co-occurrence of the tags. However, there are two main limitations in this approach. Firstly, it ignores the importance of different tags in a specific document, where some tags are more relevant to a document than others but in another document the situation can be totally different. Secondly, as described above, the tags are a set of semantic topic distributions, which are learned from plain text, and so the correlations should be modeled from the semantic level, while only considering the co-occurrences is not enough.

In this paper, we propose a novel CTL model based on the topic model to learn the correlations among the tags. In the CTL model, participation vectors of the observed tags, which take advantage of both the text information and the tags, for a semi-structured corpus are used to learn the correlations. For inference, an effective inference method is devised to learn the model parameters. The outputs of the CTL model are the tags' correlation matrix and the latent topics for documents, which are learned by utilizing the learned tag correlations. After learning the CTL, we can obtain a correlational graph which shows the relationships among the tags by ranking the correlational values. In experiments, we trained the proposed model on the DBLP corpus, where we treated authors as tags and we can visualize the correlational graph among the authors. Also, for one special author, there is a ranking list to show the relevant authors not only from the co-author information but also from whether they have similar research interests. We also apply the CTL model to the document retrieval and classification tasks on the Wikipedia corpus and the results show that the CTL model outperforms the state-of-the-art baselines.

2 RELATED WORKS

To date, many models are proposed for document modeling via different approaches such as undirected graphical models [24, 20, 13, 26, 25] or directed graphical models. As directed graphical models, topic models [11, 3, 1, 2, 4, 10] have been found to play an important role in analyzing unstructured texts. These models have been applied to many text mining areas, including information retrieval [28], document classification [6], and so on. However, most of these undirected and directed graphical models just consider the unstructured text with the bag-of-word assumption.

More and more text mining tasks are emerging in real-world applications to handle the semi-structured corpora, such as document classification described in [5, 16]. Based on the topic model, many methods have been proposed to deal with the semi-structured corpora, such as the author topic model [18], labeled-LDA [21], DMR [19], Tag-Weighted Topic Model (TWTM) [15], Tag-Weighted Dirichlet Allocation (TWDA) [14], partially LDA [22], TMBP [9], cFTM [8], statistical topic models [23], and so on. Most of the models take advantage of some given meta data (e.g., tags, labels, or contextual information) in a document with different assumptions. For example, the author topic model defines the distributions of the authors over the latent topics and the authors are assumed to be independent under a Dirichlet prior. In the labeled-LDA and partially LDA, the labels are defined as a set of distributions over the words from a vocabulary. For the TWTM and TWDA, a weight vector is used to generate the topic distribution of a document with the given tags. The DMR model is a Dirichlet-multinomial regression topic model which defines a log-linear prior on the document-topic distributions. In [23], Timothy et al. investigate a class of generative topic models for multi-label documents that associate individual word tokens with different labels, where the dependency-LDA is proposed to model the relations among the labels and words. Some of the aforementioned models can obtain the topic distribution of the tags, which can be used to measure the distance between the tags. However, they fail to directly model the correlations among tags.

3 THE CTL MODEL

In this section, we will mathematically define the Correlated Tag Learning (CTL) model, and discuss the learning and inference methods.

We use the following terminologies and notations to describe a corpus where each document is associated with a set of tags, which we call the semi-structured corpus.

Semi-Structured Corpus As a collection of M documents, we define the corpus $D = \{(\mathbf{w}^1, \mathbf{t}^1), \dots, (\mathbf{w}^M, \mathbf{t}^M)\}$, where each 2-tuple $(\mathbf{w}^d, \mathbf{t}^d)$ denotes a document with its tag vector. Let $\mathbf{w}^d = (w_1^d, \dots, w_N^d)$ denote the vector of N words associated with document d. Let $\mathbf{t}^d = (t_1^d, \dots, t_L^d)$ represent the tag vector, each element of which is a binary indicator for a tag, with L as the number of all the tags in the corpus D.

Tag Matrix Here $\mathbf{t}^d$ is expanded to a $l^d \times L$ tag matrix T^d, where l^d is the number of tags in document d for the convenience of the inference. For each $i \in \{1, \dots, l^d\}$, $T_{i\cdot}^d$ is a binary vector, where $T_{ij}^d = 1$ if

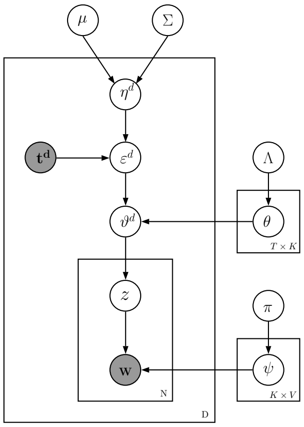


Figure 1: The graphical model of the CTL model, where each node denotes a random variable, a shaded node represents an observed variable, and edges indicate possible dependencies.

and only if the i-th tag in the document d is the j-th tag in the tag set of the corpus D.

Topic Proportions Each document is associated with a set of topic proportions ϑ. For a document d, ϑ^d is a multinomial distribution over topics and it reflects the probabilities of the words in document d drawing from latent topics.

3.1 The Model

The proposed CTL model is a hierarchical Bayesian model based on the topic model with assumptions that each document in a corpus is modeled by an underlying set of latent topics and that each topic defines a multinomial distribution over words. Besides, the CTL model assumes that the topic distribution of each document is determined by the given tags with a set of participation values. With the participation values in a participation vector, the CTL can model the topic proportions of each document as the product between the participation vector and the topic distributions of each tags in the document.

In this paper, we use ϑ_d to denote the topic distribution of the document d, as shown in Figure 1. Let θ represent a $T \times K$ matrix, where K is the number of the latent topics and each row in θ describes the distribution of one tag belonging to the latent topics. Let ψ represent a $K \times V$ distribution matrix, where each row is a distribution vector of one topic over words and V is the number of words in the dictionary of D.

3.1.1 Participation Vectors

ε^d, as shown in Figure 1, denotes the participation vector of the given tags in the document d. In the TWTM [15] and TWDA [14] models, it is called a weight vector which follows a Dirichlet distribution. As discussed above, under a Dirichlet distribution, the components of the participation vector are nearly independent, leading to a strong and unrealistic assumption that the presence of one observed tag is not correlated to the presence of another one. In order to overcome this assumption, we use a flexible logistic normal distribution to model the observed tags. As shown in Figure 1, Σ is the covariance matrix of the logistic normal distribution, μ is the expected value vector of the random variables, and η^d is a L-dimensional row vector that follows the normal distribution with Σ as the covariance and μ the mean. So the participation vector is defined as follows:

$$\varepsilon^d = \exp\{T^d \times (\eta^d)^{\mathrm{T}}\},$$

where $(\eta^d)^{\mathrm{T}}$ is the transpose of η^d, and ε^d is a $l^d \times 1$ column vector associated with the document d. Note that ε^d does not satisfy $\sum_i \varepsilon_i^d = 1$, hence we call it a participation vector instead of a weight vector.

With the participation vector, instead of a Dirichlet distribution, we use a logistic normal distribution to model the topic distribution of the document d:

$$\vartheta^d = \frac{(\varepsilon^d)^T \times T^d \times \theta}{\sum_i ((\varepsilon^d)^T \times T^d \times \theta)_i},$$

where $(\cdot)_i$ denotes the i-th entry in a vector and ϑ^d, the multinomial topic proportions of the document d, satisfies $\sum_i \vartheta_i^d = 1$. With the multinomial topic proportions ϑ obtained by a participation vector and a set of the observed tags in a document, we can generate each word for the document in a similar way to the topic model.

Thus, the CTL model assumes that a corpus with M documents arises from the following generative process:

1. For each topic $k \in \{1, \ldots, K\}$, draw $\psi_k \sim \mathrm{Dir}(\pi)$, where $\mathrm{Dir}(\cdot)$ denotes a Dirichlet distribution and π is a V-dimensional vector of hyperparameters.

2. For each tag $t \in \{1, \ldots, L\}$, draw $\theta_t \sim \mathrm{Dir}(\Lambda)$, where Λ is a K dimensional prior vector of θ.

3. For each document d:

   (a) Draw $\eta^d \sim \mathcal{N}(\mu, \Sigma)$ where $\mathcal{N}(\cdot, \cdot)$ denotes a multivariate normal distribution.

   (b) Generate T^d by $\mathbf{t}^d$.

   (c) Generate $\varepsilon^d = \exp\{T^d \times (\eta^d)^{\mathrm{T}}\}$.

   (d) Generate $\vartheta^d = \frac{(\varepsilon^d)^T \times T^d \times \theta}{\sum_i ((\varepsilon^d)^T \times T^d \times \theta)_i}$.

   (e) For each word w_{dn}:

i. Draw z_{dn} ~Mult(ϑ^d) where Mult($\cdot$) denotes a multinomial distribution.

ii. Draw w_{dn} ~Mult($\psi_{z_{dn}}$).

The CTL model is different from the TWTM model [15] where the weight vector in a document for the given tags is drawn from a Dirichlet distribution. The Dirichlet distribution is computationally convenient but it has a nearly independent assumption among the components of the weight vector. Differently, entries in the participation vector of the observed tags is highly correlated as we described above.

The covariance matrix Σ induces the dependencies between the components of the participation vector, and allows a general pattern of variability between the components. Using the covariance matrix of the logistic normal distribution, we can capture the correlated relationships between the given tags associated with each document.

3.2 Variational Inference

The logistic normal distribution used here brings not only the capacity to model the correlations among tags but also a challenge for the posterior inference procedure since it is not a conjugate prior for the multinomial distribution. We present a variational expectation-maximization (EM) algorithm [12, 27] for the inference. In the variational EM algorithm, the E-step approximates the posterior by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior distribution. This method casts the inference problem as an optimization problem to approximate the posterior distribution of this latent model and some study in [2] shows that minimizing the KL divergence is equivalent to maximizing the evidence lower bound (ELBO) denoted by $\mathcal{L}(\cdot)$.

For the CTL model, the ELBO can be derived by using Jensen's inequality:

$$\begin{aligned}\mathcal{L}(\cdot) = &\sum_d^D E_q[\log p(\eta^d|\mu, \Sigma)] + \sum_d^D \sum_n^N E_q[\log p(z_n|\vartheta^d)] \\ &+ \sum_d^D \sum_n^N E_q[\log p(w_n|\psi, z_n)] + \sum_i^L E_q[\log p(\theta_i|\Lambda)] \\ &+ H(q), \qquad (1)\end{aligned}$$

where $q(\cdot)$ denotes a variational distribution of the latent variables, $E_q[\cdot]$ denotes the expectation with respect to q, and $H(q)$ is the entropy of the variational distribution whose definition is:

$$\begin{aligned}H(q) = &-\sum_d^D E_q[\log q(\eta^d)] - \sum_d^D \sum_n^N E_q[\log q(z_n)] \\ &- \sum_i^L E_q[\log q(\theta_i)].\end{aligned}$$

For the variational distribution $q(\cdot)$, we choose a fully factorized distribution where all the variables are assumed to be independent:

$$q(\eta, z, \theta|u, \sigma^2, \gamma, \lambda) = \prod_i^L \mathrm{Dir}(\theta_i|\lambda_i) \prod_d^D \left(\mathcal{N}(\eta^d|u, \sigma^2) \prod_n^N \mathrm{Mult}(z_n|\gamma_n)\right),$$

where λ in the Dirichlet distribution, γ in the multinomial distribution, and (u, σ^2) in the Gaussian distribution are the variational parameters.

Before discussing the optimization procedure, we describe how to compute the ELBO in Eq. (1). In the CTL model, the key inferential problem that we need to solve is to compute the second term in Eq. (1), which is the expected logorithm of a topic assignment subject to a normalized multinomial parameter and can be computed as

$$\begin{aligned}E_q[\log p(z_n|\vartheta^d)] &= E_q\left[\log p\left(z_n|\frac{(\varepsilon^d)^T \times T^d \times \theta}{\sum_i((\varepsilon^d)^T \times T^d \times \theta)_i}\right)\right] \\ &= \sum_k^K \gamma_{nk} E_q\left[\log\left(\frac{(\varepsilon^d)^T \times T^d \times \theta}{\sum_i((\varepsilon^d)^T \times T^d \times \theta)_i}\right)_k\right],\end{aligned}$$

where γ_{nk} denotes the probability of the k-th topic assigned to the n-th word. We see that computing the CTL's ELBO relies on the calculation of the expected normalized topic distribution of a document, which can be computed as

$$\begin{aligned}&E_q\left[\log\left(\frac{(\varepsilon^d)^T \times T^d \times \theta}{\sum_i((\varepsilon^d)^T \times T^d \times \theta)_i}\right)_k\right] \\ =&E_q\left[\log((\varepsilon^d)^T \times T^d \times \theta)_k\right] - E_q\left[\log(\sum_i((\varepsilon^d)^T \times T^d \times \theta)_i)\right] \\ =&E_q\left[\log\sum_i^{l^d} \varepsilon_i^d \theta_k^{(i)}\right] - E_q\left[\log\sum_i^{l^d} \varepsilon_i^d\right] \\ =&E_q\left[\log\sum_i^{l^d} \exp\{\eta_{(i)}^d\}\theta_k^{(i)}\right] - E_q\left[\log\sum_i^{l^d} \exp\{\eta_{(i)}^d\}\right],\end{aligned}$$

where $\theta^{(i)}$ denotes the vector of the topic distributions for the i-th tags in the document d corresponding to a row in θ and $\eta_{(i)}^d$ is the i-th entry of $T^d \times (\eta^d)^{\mathrm{T}}$.

By following the correlated topic model [1], the above two expectations can be computed approximately with Taylor expansions, respectively:

$$E_q[\log\sum_i^{l^d} \exp\{\eta_{(i)}^d\}\theta_k^{(i)}] \approx \log\alpha + \frac{1}{\alpha}\sum_i^{l^d} E_q[\exp\{\eta_{(i)}^d\}]E_q[\theta_k^{(i)}] - 1$$

and

$$E_q[\log \sum_i^{l^d} \exp\{\eta_{(i)}^d\}] \approx \log \beta + \frac{1}{\beta} \sum_i^{l^d} E_q[\exp\{\eta_{(i)}^d\}] - 1,$$

where we introduce two new variational parameter α and β. Note that $E_q[\exp\{\eta_{(i)}^d\}]$ is the mean of a log-normal distribution and equals $\exp\{u_{(i)} + \sigma_{(i)}^2/2\}$. The expectation of a Dirichlet random variable, $E_q[\theta_k^{(i)}]$, is equal to $[\lambda_k / \sum_j^K \lambda_j]_{(i)}$. Thus, for a document d, we have

$$\begin{aligned} &\sum_n^N E_q[\log p(z_n|\vartheta^d)] \\ &\approx \sum_n^N \sum_k^K \gamma_{nk} \Bigg(\log \alpha + \frac{1}{\alpha} \sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)} \\ &\qquad - \log \beta - \frac{1}{\beta} \sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \Bigg). \end{aligned}$$

Thus, we can use the block coordinate-ascent variational inference to maximize Eq. (1) with respect to variational parameters including σ^2, u, γ, λ, α, and β.

We first maximize $\mathcal{L}(\cdot)$ with respect to σ^2 for the document d with the objective function formulated as

$$\begin{aligned} \mathcal{L}(\sigma^2) &= -\frac{1}{2} \mathrm{tr}\big(\mathrm{diag}(\sigma^2)\Sigma^{-1}\big) + \sum_i^L \frac{1}{2} \log \sigma_{(i)}^2 \\ &+ \sum_n^N \sum_k^K \frac{\gamma_{nk}}{\alpha} \Bigg(\sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)} \Bigg) \\ &- \frac{N}{\beta} \sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\}, \end{aligned} \tag{2}$$

where tr($\cdot$) denotes the trace of a square matrix and diag($\cdot$) converts a vector to a diagonal matrix. Obviously the problem with respect to σ has no analytic solution and we solve it via the Newton's method with gradient computed as

$$\begin{aligned} \mathcal{L}^{'}(\sigma_i^2) =& \frac{1}{2} \sum_n^N \sum_k^K \frac{\gamma_{nk}}{\alpha} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)} \\ &- \frac{N}{2\beta} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} + \frac{1}{2\sigma_{(i)}^2} - \frac{1}{2} \Sigma_{ii}^{-1}, \end{aligned} \tag{3}$$

where the subscript $(i) \in (1, \cdots, l^d)$ indicates the i-th tag in a specific document d.

The objective function with respect to u is formulated as

$$\begin{aligned} \mathcal{L}(u) &= -\frac{1}{2}(u-\mu)^T \Sigma^{-1}(u-\mu) - \frac{N}{\beta} \sum_i^{l^d} \exp\{u_{(i)} + \frac{\sigma_{(i)}^2}{2}\} \\ &+ \sum_n^N \sum_k^K \frac{\gamma_{nk}}{\alpha} (\sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)}). \end{aligned} \tag{4}$$

We use the conjugate gradient algorithm to solve this problem, where the derivative is computed as

$$\begin{aligned} \mathcal{L}^{'}(u) =& \sum_n^N \sum_k^K \frac{\gamma_{nk}}{\alpha} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)} \\ &- \frac{N}{\beta} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} - \Sigma^{-1}(u-\mu). \end{aligned} \tag{5}$$

We maximize Eq. (1) with respect to γ_{nk} to find the maximizer as

$$\begin{aligned} \gamma_{nk} \propto& \psi_{k,v^{w_n}} \exp\Bigg\{ \frac{1}{\alpha} (\sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)}) \\ &+ \log \alpha \Bigg\}, \end{aligned} \tag{6}$$

where v^{w_n} denotes the index of w_n in the dictionary.

For the variational parameter λ, the objective function is formulated as

$$\begin{aligned} \mathcal{L}(\lambda) =& \sum_k^K (\Lambda_k - 1)(\Psi(\lambda_k) - \Psi(\sum_j^K \lambda_j)) - \log \Gamma(\sum_j^K \lambda_j) \\ &+ \sum_n^N \sum_k^K \frac{\gamma_{nk}}{\alpha} \Bigg(\sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)} \Bigg) \\ &+ \sum_k^K \log \Gamma(\lambda_k) + \sum_k^K (\lambda_k - 1)(\Psi(\lambda_k) - \Psi(\sum_j^K \lambda_j)). \end{aligned} \tag{7}$$

We use the gradient descent method to solve it, where the derivative with respect to λ_k is:

$$\begin{aligned} \mathcal{L}^{'}(\lambda_k) =& \sum_n^N \frac{\gamma_{nk}(\sum_j^K \lambda_j - \lambda_k)}{\alpha(\sum_j^K \lambda_j)^2} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \\ &+ (\Lambda_k - \lambda_k)(\Psi^{'}(\lambda_k) - \Psi^{'}(\sum_j^K \lambda_j)). \end{aligned} \tag{8}$$

For α and β, the optimal solutions can easily be found as

$$\alpha \propto \frac{\sum_n^N \sum_k^K \gamma_{nk} \Big(\sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\} \left[\frac{\lambda_k}{\sum_j^K \lambda_j} \right]_{(i)} \Big)}{\sum_n^N \sum_k^K \gamma_{nk}} \tag{9}$$

$$\beta \propto \sum_i^{l^d} \exp\{u_{(i)} + \sigma_{(i)}^2/2\}. \tag{10}$$

In the E-Step of the variational EM algorithm, we iteratively update the variational parameters including σ^2 , u, γ, λ, α and β.

3.3 Parameter Estimation

The parameters of the CTL model include Σ, μ, ψ and Λ. In the M-step, given the semi-structured corpus, we can estimate the parameters by maximizing a lower-bound of the log-likelihood based on the variational

E-step. The update rules for Σ, μ and ψ can easily be obtained:

$$\mu \propto \frac{1}{D}\sum_{d}^{D} u_d, \tag{11}$$

$$\Sigma \propto \frac{1}{D}\sum_{d}^{D}\Big(I\sigma_d^2 + (u_d-\mu)(u_d-\mu)^T\Big), \tag{12}$$

$$\psi_{kj} \propto \sum_{d}^{D}\sum_{n}^{N}\gamma_{nk}(w^d)_n^j, \tag{13}$$

where $(w^d)_n^j$ is the count for the n-th word in the document d. For the Dirichlet parameter Λ, its objective function is formulated as

$$\begin{aligned}\mathcal{L}(\Lambda) = \sum_{l}^{L}\Bigg(&\log\Gamma(\sum_{j}^{K}\Lambda_j) - \sum_{i}^{K}\log\Gamma(\Lambda_i)\\ &+\sum_{i}^{K}(\Lambda_i-1)\Big(\Psi(\lambda_i^l) - \Psi(\sum_{j}^{K}\lambda_j^l)\Big)\Bigg).\end{aligned} \tag{14}$$

The derivative with respect to Λ_i is computed as

$$\mathcal{L}^{'}(\Lambda_i) = L\Big(\Psi(\sum_{j}^{K}\Lambda_j) - \Psi(\lambda_i)\Big) + \sum_{l}^{L}\Big(\Psi(\lambda_i^l) - \Psi(\sum_{j}^{K}\lambda_j^l)\Big). \tag{15}$$

We can use the linear-time Newton-Raphson algorithm to estimate Λ.

4 DISCUSSION

The proposed CTL model can capture the correlations among the tags in a semi-structured corpus, not just only by considering the co-occurrences of the tags. The CTL model presents the participation vector for each document to estimate the correlations of the tags, and the participation vector is learned by the text information with the basic assumption on latent topics. In other words, the co-occurrence vector is binary, which means that one tag is present or absent, while the participation vector is non-binary and the values in participation vector denote the importance of the tags.

Actually, we can train the CTL model by only considering the co-occurrence information of tags in each document. In this case, different tags have equal importance in a document d and hence η^d is observed for each document, where η_j^d is set to 1 if and only if the document d has the j-th tag. So the CTL model will consist of two parts, as shown in Figure 2. The left figure in Figure 2 is the first part containing Σ, μ, η^d, ξ^d and $\mathbf{t}^d$, where η^d, ξ^d and $\mathbf{t}^d$ are the observed variables. We can use the traditional maximum likelihood estimation to learn the correlation matrix Σ with the D samples:

$$\mu = \frac{1}{D}\sum_{d}^{D}\mathbf{t}^d, \quad \Sigma = \frac{1}{D}\sum_{d}^{D}(\mathbf{t}^d-\mu)(\mathbf{t}^d-\mu)^{\mathrm{T}}.$$

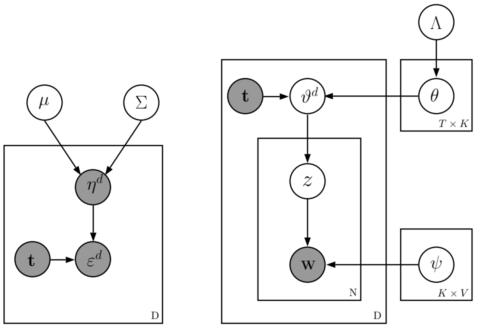


Figure 2: The two parts of the CTL model will be degenerated if η^d becomes equal to $\mathbf{t}^d$, which means that all the tags have the same effect on the document.

The second part shown in the right figure of Figure 2 means that all the tags have equal impacts on the topic distribution ϑ^d. We can see that the second part is a variant of the author topic model described in [18]. Thus, we can use the variational inference process to compute the new ELBO bound as:

$$\begin{aligned}\mathcal{L}_{new} = &\sum_{d}^{D}E_q[\log p(\mathbf{t}^d|\mu,\Sigma)] + \sum_{d}^{D}\sum_{n}^{N}E_q[\log p(z_n|\vartheta^d)]\\ &+\sum_{d}^{D}\sum_{n}^{N}E_q[\log p(w_n|\psi,z_n)] + \sum_{i}^{L}E_q[\log p(\theta_i|\Lambda)]\\ &-\sum_{d}^{D}E_q[\log q(\mathbf{t}^d)] - \sum_{d}^{D}\sum_{n}^{N}E_q[\log q(z_n)]\\ &-\sum_{i}^{L}E_q[\log q(\theta_i)],\end{aligned}$$

where $\sum_d^D E_q[\log p(\mathbf{t}^d|\mu,\Sigma)]$ and $\sum_d^D E_q[\log q(\mathbf{t}^d)]$ are fixed, $\sum_d^D\sum_n^N E_q[\log p(z_n|\vartheta^d)]$ does not involve σ^2 and u since η^d and ξ^d are known. In this case, $\mathcal{L}_{new} < \mathcal{L}$, which means the new lower bound $\mathcal{L}_{new}$ is lower than the former one when convergence. Thus, treating the tags equally will not be a good choice.

Compared with the tag-weighted topic model [15], we would obtain document embeddings with better quality when the tags in the corpus are highly correlated. Thus, we will study the CTL model under this setting in our experiments.

5 EXPERIMENTS

In this section, we will present the performance of the proposed CTL model on document modeling, document classification, and document retrieval, respectively.

5.1 Experimental Settings

We used two semi-structured corpora to evaluate the CTL model. The first corpus is the Digital Bibliography and Library Project (DBLP),[1] which is a collection of bibliographic information of technical papers published in major computer science journals and conferences. We use the authors as the tags and removed the authors that occur in fewer than 5 papers. We use a subset of the DBLP that contains abstracts of $D = 40,108$ papers with $72,748$ words by removing stop words and $L = 6,348$ unique tags. The second corpus is from Wikipedia.[2] The Wikipedia corpus we used contains $43,217$ articles. The size of the vocabulary is $22,344$ by removing stop words. We use the category information, which is located at the bottom of each article and provided by the MediaWiki software, of articles as the tags, and in total there are $2,900$ tags. Moreover, each article belongs to different portals which can be viewed as the class label and all the articles used in the experiments belong to 20 classes, such as arts, sports, history, biography, education and so on.

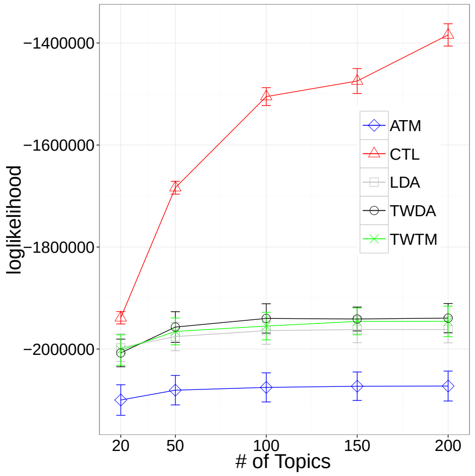
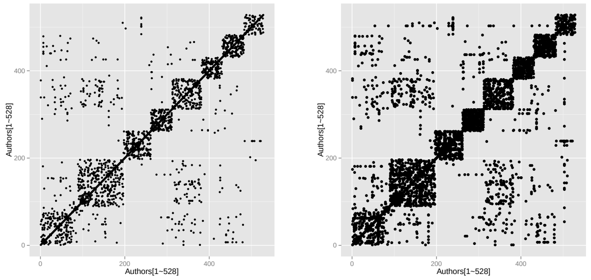


Figure 3: The 5-fold cross-validated held-out log-likelihood of different models on the Wikipedia corpus with different number of topics.

5.2 Experiments on Document Modeling

To demonstrate the performance of the different models on document modeling, we computed the log-likelihood of the held-out data given a model estimated from the remaining data by using five-fold cross validation, implying that 80% documents are for training and the remaining 20% for testing. We compared the CTL model with the Author Topic Model (ATM), TWTM, TWDA, and LDA by varying the number of latent topics. Since the LDA could not handle the tag information directly, we treated the given tags as the word features and added them into the document as the input for the LDA model.



(a) penalty factor $p = 0.25$ (b) penalty factor $p = 0.5$

Figure 4: The scatter plots of the selected 528 authors on the DBLP corpus, where a point is drawn if the corresponding two authors are neighbors.

Figure 3 shows the average log-likelihood for each model on the held-out set. The results demonstrate that the CTL model has much better performance than other baselines. One possible reason is that the tags contained in Wikipedia corpus are highly correlated and with the help of the logistic normal distribution, the CTL model can obtain a more reasonable and effective participation vector to form the topic distribution for each document.

5.3 Analysis on Tag Graph

The covariance of the logistic normal distribution for the participation vector can be used to visualize the relations among tags. Thus, we use the covariance matrix to form a tag graph, where the nodes represent the tags appeared in the corpus and the edges denote the relations between tags. To construct the tag graph, we use the method introduced in [17] for neighborhood selection based on the Lasso. As described in [17], the neighborhood selection with the Lasso is used to estimate the conditional dependency separately for each node in the graph. In the CTL model, for a document d, η^d follows a normal distribution with mean u and covariance Σ. Thus, $\{\eta^d\}$ are treated as independent observations sampled from the normal distribution $\mathcal{N}(\mu, \Sigma)$, which are used to estimate the neighborhood based on [17].

We use the DBLP corpus for the experiment. For the convenience of display, we select 528 authors to illustrate the correlated connections among them by drawing an edge if the corresponding two authors are neighbors with different penalty factor $p = 0.25$ and

[1] `http://www.informatik.uni-trier.de/~ley/db/`

[2] http://www.wikipedia.org/

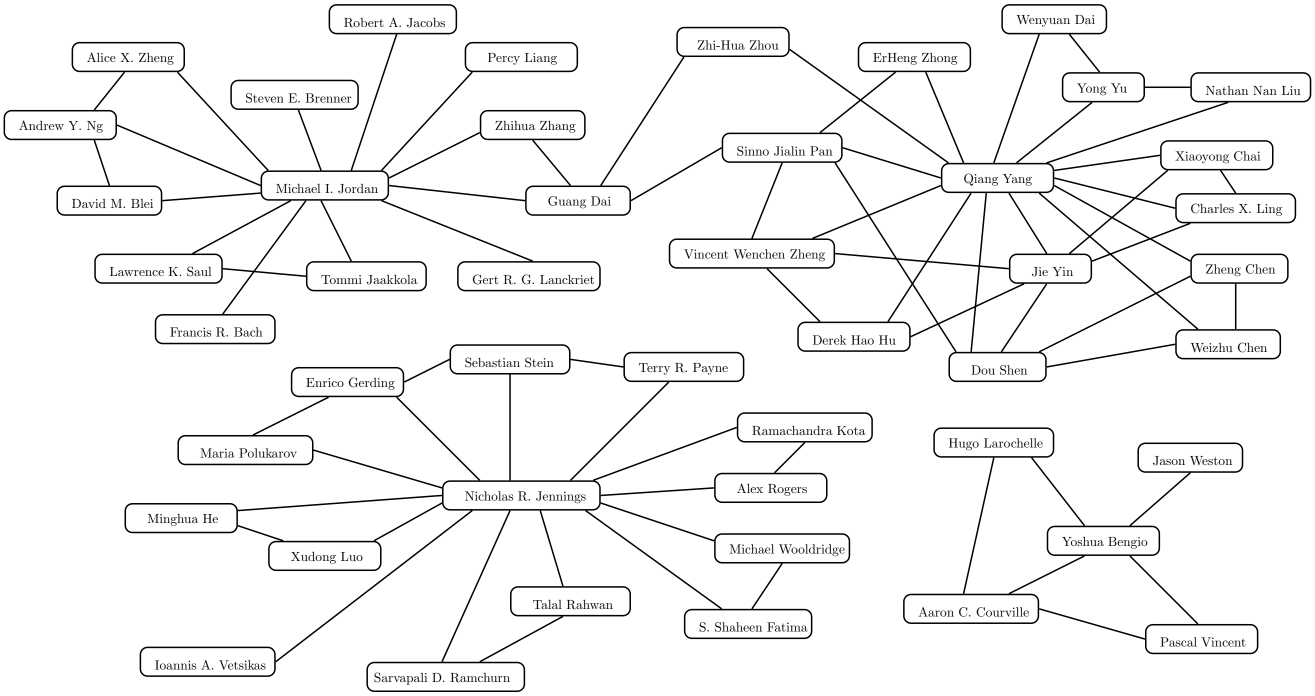


Figure 5: A subset of the author graph learned from 40,108 abstracts of the DBLP. The edges between authors are computed by the neighborhood selection method [17] based on the Lasso.

Table 1: The ranking list of top correlated authors with eight authors on the DBLP corpus.

Michael I. Jordan	Alice X. Zheng(0.157837), Francis R. Bach(0.116478), Gert R. G. Lanckriet(0.107671), David M. Blei(0.103741), Steven E. Brenner(0.092675), Zhihua Zhang(0.090492), Percy Liang(0.089226), Robert A. Jacobs(0.085318), Tommi Jaakkola (0.082044), Guang Dai(0.058045), Lawrence K. Saul(0.057545), Martin J. Wainwright(0.045444), Nebojsa Jojic(0.043731), David A. Patterson(0.041441), Tamar Flash(0.035152), Erik B. Sudderth(0.033563), Andrew Y. Ng(0.013234)
Yoshua Bengio	Pascal Vincent(0.410208), Hugo Larochelle(0.265349), Aaron C. Courville(0.132399), Jason Weston(0.049443)
Qiang Yang	Dou Shen(0.149414), Derek Hao Hu(0.144670), Sinno Jialin Pan(0.134137), Xiaoyong Chai(0.115358), Nathan Nan Liu (0.0.113420), ErHeng Zhong(0.107736), Weizhu Chen(0.104334), Yong Yu(0.091473), Charles X. Ling(0.081368), Jie Yin (0.077154), Wenyuan Dai(0.071907), Vincent Wenchen Zheng(0.062750), Zheng Chen(0.060292), Zhi-Hua Zhou(0.057265)
Nicholas R. Jennings	Alex Rogers(0.243331), Ramachandra Kota(0.142240), Maria Polukarov(0.137646), Talal Rahwan(0.134888), Ioannis A. Vetsikas(0.126621), S. Shaheen Fatima(0.121799), Sebastian Stein(0.117208), Enrico Gerding(0.027282), Minghua He(0.106050), Xudong Luo(0.086897), Sarvapali D. Ramchurn(0.071510), Terry R. Payne(0.068713), Michael Wooldridge(0.065824)
Micha Sharir	Pankaj K. Agarwal(0.273844), Emo Welzl(0.170932), Natan Rubin(0.139646), Jnos Pach(0.110192), Haim Kaplan (0.106137), Vladlen Koltun(0.104685), Boris Aronov(0.102123), Shakhar Smorodinsky(0.088530), Esther Ezra(0.083021), Dan Halperin(0.074751, Rom Pinchasi(0.056282), Bernard Chazelle(0.044011), Jir Matousek(0.027263)
Jiawei Han	Xiaoxin Yin(0.196956), Deng Cai(0.103920), Guozhu Dong(0.101735), V. S. Lakshmanan(0.100016), Xin Jin(0.098451), Charu C. Aggarwal(0.098256), Anthony K. H. Tung(0.092360), Jianyong Wang(0.087017), Hongjun Lu(0.072772), ChengXiang Zhai(0.048570), Jiong Yang(0.042489), Philip S. Yu(0.036918), Ke Wang(0.032003)
Jennifer Rexford	David Walker(0.247776), Mung Chiang(0.162803), Eric Keller(0.152502), Renata Teixeira(0.141469), Minlan Yu(0.123096), Nick Feamster(0.115787), Albert G. Greenberg(0.111072), Aman Shaikh(0.093805), Matthew Caesar(0.049353), Michael J. Freedman(0.039058), Kang G. Shin(0.031615)
Franco Zambonelli	Marco Mamei(0.372887), Letizia Leonardi(0.213969), Giacomo Cabri(0.195711), Gabriella Castelli(0.188215), Nicola Bicocchi(0.080601), Andrea Omicini(0.054696), Robert Tolksdorf(0.040949), Sara Montagna(0.020905), Matthias Baumgarten(0.012859), Alberto Rosi(0.012646)

0.5, respectively. After using the spectral clustering based on the conditional dependency obtained by the method in [17], as shown in Figure 4, we clearly see that the authors cluster together to different groups, where the authors in a group may have similar research interests.

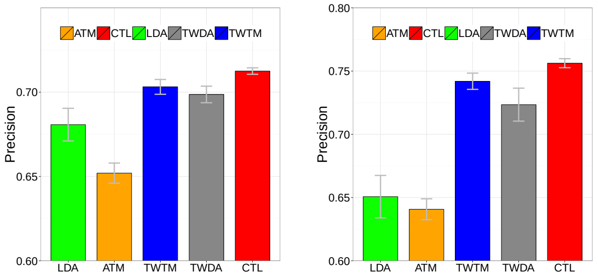
We plot the author graph to show the correlations among the 518 authors and in Figure 5, a part of the graph is shown, where the nodes represent authors and the edges denote the correlations between the authors. In this graph, we can find some interesting insight. For example, two authors ‘Zhi-Hua Zhou’ and ‘Guang Dai’, who did not coauthor any paper, have a connection between them since they have similar research interests, which shows an advantage of the CTL model over only using the coauthorship information that it can find meaningful relations at the semantic level.

The CTL model can give a ranking list based on how the authors are correlated. In Table 1, we pick several authors from the 528 authors, and for each selected author, we rank other authors according to their correlational values, which can be obtained by the process of neighborhood selection (see [17]). Based on the results, we can easily see how close the two authors are in the research and answer interesting questions including that whether researcher A has more similar interests to researcher B than to researcher C.

5.4 Document Classification and Retrieval

In this section, we conduct experiments on document classification and retrieval tasks.

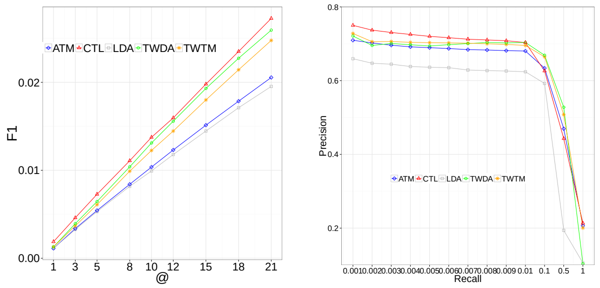
We first test the classification performance by comparing the performance of the LDA, ATM, TWTM, TWDA, and the proposed CTL model with the number of topics as 50 and 100, respectively. The LIBSVM [7] with the Gaussian kernel and default parameters is used as the classifier. In experiments, we use a subset of the Wikipedia corpus which contains 14, 400 documents belonging to 20 classes. We reported in Figure 6 the precision of different methods on the Wikipedia corpus by using the five-fold cross validation. According to Figure 6, the performance of the CTL model is significantly better than that of other baseline methods. One possible reason is that the CTL model can learn a better topic distribution for each document than others.



(a) $K = 50$ (b) $K = 100$

Figure 6: Classification results on the Wikipedia corpus for the LDA, ATM, TWTM, TWDA and CTL models with five-fold cross validation.

Moreover, we use the Wikipedia corpus to evaluate the performance on the document retrieval task. In this experiment, each document is represented by the vector of topic distribution generated by different models with the topic number as $K = 100$. The Wikipedia corpus used here is just the data set used in the above classification experiment. We randomly sample 12, 400 documents for training and the rest for testing. For each query, documents in the database were ranked using the cosine distance as the similarity metric. For evaluation, we check whether a retrieved document has the same class label as the query document to decide whether the retrieved document is relevant to the query document. Figure 7 shows the F1 scores and the precision-recall curves of the LDA, ATM, TWTM, TWDA and CTL models. The experimental results demonstrate the superiority of the embedding learned by the CTL model for the document retrieval task.



(a) F1-Score (b) Precision-Recall curves

Figure 7: F1-score and precision-recall curves for document retrieval on the Wikipedia corpus for the LDA, ATM, TWTM, TWDA, and CTL models with $K = 100$.

6 CONCLUSION

In this paper, we propose the CTL model, a statistical model of semi-structured corpora, based on the topic model to discover highly correlational relationships among the tags observed in the semi-structured corpus. Besides, the experimental results demonstrated that this method can model semi-structured corpora better than the state-of-the-art models when the tags are highly correlated.

In our future study, we will apply the CTL model to more text applications. Another possible direction is to devise parallel algorithms for the CTL model to further improve its efficiency.

ACKNOWLEDGEMENT

We thank the support of China National 973 project 2014CB340304 and Hong Kong CERG projects 16211214 and 16209715.

References

[1] David M. Blei and John D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[2] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Jordan L. Boyd-Graber and David M. Blei. Syntactic topic models. *CoRR*, abs/1002.4665, 2010.

[5] Andrej Bratko and Bogdan Filipic. Exploiting structural information for semi-structured document categorization. *Information Processing and Management*, 42(3):679 – 694, 2006.

[6] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. In *CIKM*, pages 911–920, 2008.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[8] Xu Chen, Mingyuan Zhou, and Lawrence Carin. The contextual focused topic model. In *KDD*, pages 96–104, 2012.

[9] Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *KDD*, pages 1271–1279, 2011.

[10] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online learning for latent Dirichlet allocation. In *NIPS*, pages 856–864, 2010.

[11] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[12] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[13] Hugo Larochelle and Stanislas Lauly. A neural autoregressive topic model. In *NIPS*, pages 2717–2725, 2012.

[14] Shuangyin Li, Guan Huang, Ruiyang Tan, and Rong Pan. Tag-weighted Dirichlet allocation. In *ICDM*, pages 438–447, 2013.

[15] Shuangyin Li, Jiefei Li, and Rong Pan. Tag-weighted topic model for mining semi-structured documents. In *IJCAI*, 2013.

[16] Pierre-Francois Marteau, Gildas Ménier, and Eugen Popovici. Weighted naive Bayes model for semi-structured document categorization. *CoRR*, abs/0901.0358, 2009.

[17] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[18] Rosen-Zvi Michal, Chemudugunta Chaitanya, Griffiths Thomas, Smyth Padhraic, and Steyvers Mark. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38, 2010.

[19] David M. Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI*, pages 411–418, 2008.

[20] Srivastava Nitish, Ruslan Salakhutdinov, and Geoffrey E. Hinton. Modeling documents with a deep Boltzmann machine. In *UAI*, 2013.

[21] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256, 2009.

[22] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *SIGKDD*, pages 457–465, 2011.

[23] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208, 2012.

[24] Ruslan Salakhutdinov and Geoffrey E. Hinton. Replicated softmax: an undirected topic model. In *NIPS*, pages 1607–1614, 2009.

[25] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *ICML Workshop*, 2012.

[26] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep Boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[27] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[28] Xing Wei and W. Bruce Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.