2013 IEEE 13th International Conference on Data Mining

Tag-Weighted Dirichlet Allocation

Shuangyin Li, Guan Huang, Ruiyang Tan and Rong Pan*
School of Information Science and Technology
Sun Yat-sen University, Guangzhou, China
{lishyin@mail2., huangg6@mail2., tanry@mail2., panr@}sysu.edu.cn

***Abstract*—In the past two decades, there has been a huge amount of document data with rich tag information during the evolution of the Internet, which can be called semi-structured data. These semi-structured data contain both unstructured features (e.g., plain text) and metadata, such as tags in html files or author and venue information in research articles. It's of great interest to model such kind of data. Most previous works focused on modeling the unstructured data. Some other methods have been proposed to model the unstructured data with specific tags. To build a general model for semi-structured documents remains an important problem in terms of both model fitness and efficiency. In this paper, we propose a novel method to model the tagged documents by a so-called Tag-Weighted Dirichlet Allocation (TWDA). TWDA is a framework that leverages both the tags and words in each document to infer the topic components for the documents. This allows not only to learn the document-topic and topic-word distributions, but also to infer the tag-topic distributions for text mining (e.g., classification, clustering, and recommendations). Moreover, TWDA can automatically infer the probabilistic weights of tags for each document, that can be used to predict the tags in one document. We present an efficient variational inference method with an EM algorithm for estimating the model parameters. The experimental results show the effectiveness, efficiency and robustness of our TWDA approach by comparing it with the state-of-the-art methods on four corpora in document modeling, tags prediction and text classification.**

I. Introduction

In the evolution of the Internet, there have been massive collection of documents in many web applications. Such kinds of documents with both text data and document metadata (tags, which can be viewed as features of the corresponding document) are called the semi-structured data. How to characterize the semi-structured document data becomes an important issue addressed in many areas, such as information retrieval, artificial intelligence and data mining, etc. The tags can be more important than the text data in document mining. There are many examples: in a collection of movie set, such as Internet Movie Database (IMDB)[1], we may have an idea that a movie has a higher chance to be a comedy when it has a tag "Dick Martin", without watch it; in a collection of scientific articles, each document has a list tags(authors and keywords).

Many solutions have been proposed to deal with the semi-structured documents (e.g., SVD, LSI), and shown to be useful in document mining [7], [20], [29], [27], e.g., text classification and structural information exploiting. For document modeling, topic models have been used to be a powerful method of analyzing and modeling of document corpora, using Bayesian statistics and machine learning to discover the thematic contents of untagged documents. Topic models can discover the latent structures in documents and establish links between them, such as latent Dirichlet allocation (LDA) [5]. However, as an unsupervised method, only the words in the documents are modeled in LDA. Thus, LDA could only treat the tags as word features rather than a new kind of information for document modeling.

To model semi-structured data needs to consider the characteristics of different kinds of objects, including word, topic, document, and tag, and also the relationship among them. In this problem, topic is a kind of hidden objects, and the other three are the observations. Relative to tag, word and document are objective; tag can be either objective (e.g., author and venue information of publications) and subjective (e.g., tags in social bookmark marked by people). Similar to the topic models, we should consider binary relationships between the pairs of the objects, including topic-word and document-topic. In addition, we may consider the binary relationships, like tag-word, tag-topic, tag-document, and tag-tag. The tag-document relationship implies that we should consider the weights of the tags in each document. The tag-topic and tag-tag relationships can be more complicated, thus are difficult to model. Some earlier works consider certain tags. For example, the author-topic model in [26] considers the authorship information of the documents to be modeled. In this work, we don't limit the types and number of the tags in each document. In an extreme case, where there is no tag in any document, the new model may degenerate into LDA. On the other hand, since the tags can be created by some people, they should be relevant to topics of the documents; however, some of them may be correlated, redundant, and even noisy. Therefore, the tag-topic relationships should be general enough and we should also model the weights of the tags in each document.

In the past few years, researchers have proposed approaches to model documents with tags or labels [21], [24], [25]. Moreover, Labeled LDA [24] assumes there is no latent topics and each topic is restricted to be associated with the given labels. PLDA assumes that each topic is associated with only one label [25]. Both Labeled LDA and PLDA have implicit assumptions that the given labels should be strongly associated with the topics to be modeled or the labels are independent to each other.

In this paper, we propose a tag-weighted Dirichlet allocation (TWDA) to represent the text data and the various tags with weights to evaluate the importance of the tags, which provides a novel method to model the semi-structured documents.

It is based on latent Dirichlet allocation (LDA), and learns

*Corresponding author
[1]http://www.imdb.com

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.11

the weights among the Dirichlet prior and tags, not just among the tags. Therefore, TWDA handle not only the semi-structured documents, but also the unstructured documents. For the unstructured documents, TWDA degenerates into LDA. Moreover, in many web applications, not all the documents in the corpora have the tags. There are lots of documents without any tags or documents which have to remove all the tags after data preprocessing for denoising. Only considering the tags would not hold this case. However, TWDA can handle this complex corpora effectively and easily.

Besides, TWDA also infers the topic distributions of tags. The weights of observed tags in each document, which we infer from the data set, give us an opportunity to provide a method to rank the tags. The contribution of this paper is three-fold:

1) It provides a novel and accurate topic modeling method to model the semi-structured data, leveraging the weights among the Dirichlet prior and observed tags in a document.
2) In TWDA, weights are associated with the observed tags in a document providing a way to rank the tags. In addition, this could be used to predict latent tags in the document.
3) With TWDA, we can handle both the multi-tag documents and non-tag documents simultaneously, which is very useful to process some complicated web applications.

The rest of the paper is organized as follows. In Section II, we first analyze and discuss related works. In Section III, after introducing the notations, we present a novel topic model, and give the methods of learning and inference. In Section IV, we present the experimental results on four domains to show the performance of the proposed method in document modeling, tag predicting, text classification and the model robustness. We end the paper in Section V.

II. Related Works

Topic models provide an amalgam of ideas drawn from mathematics, computer science, and cognitive science to help users understand unstructured data. There are many topic models proposed and shown to be powerful on document analyzing, such as in [23], [14], [5], [4], [6], [10], which have been applied to many areas, including document clustering and classification [8], and information retrieval [28]. They are extended to many other topic models for different situation of applications in analyzing text data [15], [17], [30]. However, most of these models only consider the textual information and can only treat the tag information as plain text as well.

TMBP [12] and cFTM [11] propose the methods to make use of the contextual information of documents for topic modeling. TMBP is a topic model with biased propagation to leveraging contextual information, the authors and venue. TMBP needs to predefine the weights of the author and venue information on word assignment, which limits the usefulness in real applications. The method of cFTM has a very strong assumption that each word is associated with only one tag, either author or venue.

Several models have been proposed to take advantage of tags or labels, such as Labeled LDA [24], DMR [21] and PLDA [25], or modeling relationships among several variables, such as Author-Topic Model [26]. Labeled LDA [24] get the topic distribution for a document through picking out the several hyperparameter components that correspond to its labels, and draw the topic components by the new hyperparameter without inferring the topic distribution of labels. Labeled LDA does not assume the existence of any latent topics [25]. PLDA [25] provides another way of modeling the tagged text data, which assumes the generation topics assignment is limited by only one of the given tags for one word, and in the training process, PLDA assumes that each topic takes part in exactly one label, and may optionally share global label present on every document. In Author Topic Model, it obtains the topic distributions of authors, without giving the importance weights among the given authors in each document. DMR [21] is a Dirichlet-multinomial regression topic model that includes a log-linear prior on the document-topic distributions, which is an exponential function of the given features of the document. However, DMR doesn't output the tag weights either [26], which is useful for tag ranking.

TWTM [18] proposes a approach using the tags to document modeling, however, it just leverages the tags given in a document by the weight values to model the topic distribution of a document. In many real applications, when it comes to a corpora in which some documents have no tags, TWTM doesn't work. So, in our paper, we consider both the Dirichlet prior and the tags in the document by the weight values among them, which provide a novel method to integrate tag information into probabilistic topic models. The method can automatically degenerate into LDA, treating the documents without tags as unstructured data. Besides the topic discovery, the method can be used to predict tags of documents by the weight values with a better performance.

III. TWDA Model And Algorithms

In this section, we will mathematically define the tag-weighted Dirichlet allocation (TWDA), and discuss the learning and inference methods.

A. Notation

Similar to LDA [5], we formally define the following terms. Consider a semi-structured corpus, a collection of M documents. We define the corpus $D = \{(\mathbf{w}^1, \mathbf{t}^1), \ldots, (\mathbf{w}^M, \mathbf{t}^M)\}$, where each 2-tuple $(\mathbf{w}^d, \mathbf{t}^d)$ denotes a document, the bag-of-word representation $\mathbf{w}^d = (w_1^d, \ldots, w_N^d)$, $\mathbf{t}^d = (t_1^d, \ldots, t_L^d)$ is the document tag vector, each element of which being a binary tag indicator, and L is the size of the tag set in the corpus D. For the convenience of the inference in this paper, $\mathbf{t}^d$ is expanded to a $l^d \times (L+1)$ matrix T^d, where l^d is one more than the number of tags in document d (For example, if the document d has five tags, l^d is six). For each row number $i \in \{1, \ldots, l^d\}$ in T^d, $T_{i\cdot}^d$ is a binary vector, where $T_{ij}^d = 1$ if and only if the i-th tag of the document d[2] is the j-th tag of the tag set in the corpus D. Note that, we set the last dimension of the last row in T^d to 1, and the other dimensions of the last row equal to 0 for all documents. The detail of the above setting will be shown later.

[2]Note that we can sort the tags of the document d by the index of the tag set of the corpus D.

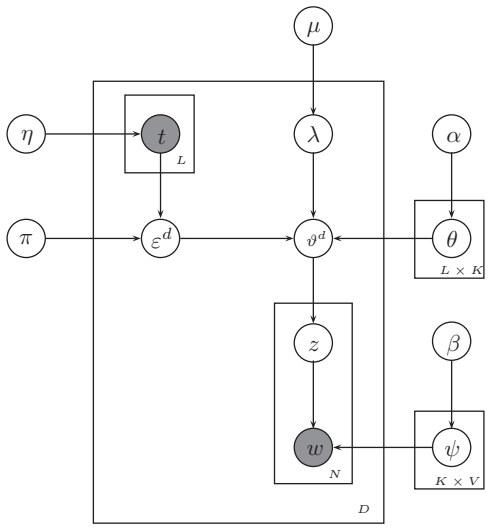


Fig. 1. Graphical model representation of TWDA, where θ is a distribution matrix of all the tags, ψ is a distribution matrix of words. ϑ^d indicates the topic proportions of each document

In this paper, we wish to find a probabilistic model for the corpus D that assigns high likelihood to the documents in the corpus and other documents alike utilizing the given tag information.

B. TWDA

TWDA is a probabilistic graphical model that describes a process for generating a semi-structured document collection. In topic models, such as LDA, we treat the words in a document as generating from a set of latent topics which is a set distributions over the vocabulary in the corpora. However, there are many tags in the semi-structured document collection and the tags have an important impact on the topic distributions in documents. In this proposed model TWDA, we try to consider the gain from tags information to the document topic modeling.

Tag-weighted Dirichlet allocation (TWDA) is built on latent Dirichlet allocation (LDA) [5], in which we treat the topic proportions for a document as a draw from a Dirichlet distribution. In tag-weighted Dirichlet allocation (TWDA), the topic distribution of one document is not only decided by a Dirichlet distribution parameter, but also by all the tags appeared in the document. The TWDA mixes the topic proportions draw from the Dirichlet distribution hyperparameter and given tags by importance or weight (tag-weighted).

The model parameters are as follows: a $K \times V$ matrix ψ (each ψ_k is a vector of words probabilities), an $L \times K$ matrix θ (each row presents the topic distribution of one tag), a Dirichlet hyperparameter μ, a Dirichlet parameter π(π is the Dirichlet parameter of weights among specified tags), and a Bernoulli prior η for model completeness. Figure 1 shows that how TWDA works in a probabilistic graphical model. In the model, we use ϑ^d to denote the topic distribution of document d and ξ^d to denote the weight vector of the tags in d as shown in Figure 1. This parameters are described in detail later.

The generative process for TWDA is given in the following procedure:

1) For each topic $k \in \{1, \ldots, K\}$, draw $\psi_k \sim$ Dir(β) , where β is a V dimensional prior vector of ψ.
2) For each tag $t \in \{1, \ldots, L\}$, draw $\theta_t \sim$ Dir(α), where α is a K dimensional prior vector of θ.
3) For each document d:
   a) Draw $\lambda \sim Dir(\mu)$.
   b) Generate T^d by $\mathbf{t}^d$.
   c) Draw $\varepsilon^d \sim Dir(T^d \times \pi)$.
   d) Generate $\vartheta^d = (\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda})$.
   e) For each word w_{di}:
      i) Draw $z_{di} \sim$Mult(ϑ^d) .
      ii) Draw $w_{di} \sim$Mult($\psi_{z_{di}}$) .

In this process, Dir($\cdot$) designates a Dirichlet distribution, Mult($\cdot$) is a multinomial distribution. Note that, L is the number of tags appeared in the corpora and K is the number of topics. π is a $(L+1) \times 1$ column vector and μ is a $K \times 1$ column vector. Both of them are Dirichlet prior. λ is a $1 \times K$ row vector which is drawn from μ. $(\varepsilon^d)^{\mathrm{T}}$ is the transpose of ε^d and ε^d is drawn from a Dirichlet prior which obtained by the matrix multiplication of $T^d \times \pi$. Clearly, the result of $T^d \times \pi$ will be a $(l^d \times 1)$ vector whose dimension is depended on the number of the observed tags in the document d. Note that, l^d is one more than the number of tags given in d as we described above. $(\varepsilon^d)^{\mathrm{T}}$ is the transpose of ε^d and ε^d is drawn from a Dirichlet prior which obtained by the matrix multiplication of $T^d \times \pi$. Clearly, the result of $T^d \times \pi$ will be a $(l^d \times 1)$ vector whose dimension is depended on the number of the observed tags in the document d.

In other words, we treat the λ as a topic distribution of one latent tag, the Dirichlet prior μ. Each document is controlled by a latent tag, that is the same idea both TWDA and Latent Dirichlet Allocation[5]. The form of $(\frac{\theta}{\lambda})$ is the augmented matrix of θ and λ, which represents that we add the vector λ to the matrix θ as the last row, so $(\frac{\theta}{\lambda})$ becomes a $(L+1) \times K$ matrix. As we show above, T^d is the matrix form of the given tags in the document d, and the last row of T^d is a binary vector, of which only the last dimension equals to 1 and the others equal 0. Here we define

$$\Theta^d = T^d \times (\frac{\theta}{\lambda}).$$

Clearly, Θ^d is a $l^d \times K$ matrix, whose last row is λ. Actually, the purpose of Θ^d is to pick out the rows corresponded to the tags appeared in d from tag-topic distribution matrix θ.

In the proposed model, the key idea of tag-weighted Dirichlet allocation is to model the topic proportions of semi-structured documents by document-special tags and text data. Different from LDA, the topic proportion of one document assumed in this paper is controlled not only by a Dirichlet prior μ, but also by all the observed tags. The way to generate the normalized topic distribution of the document d is that we mix both Dirichlet allocation and tags information through a weight vector ε^d. Thus, we obtain the topic distribution of d by

$$\vartheta^d = (\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}).$$

It is worth to note that the ε^d is draw by a Dirichlet prior π, each row of θ is draw by a Dirichlet prior α, and λ is draw by a Dirichlet prior μ, so ε^d and θ satisfy

$$\sum_{i=1}^{l^d} \varepsilon_i^d = 1, \sum_{k=1}^{K} \theta_{lk} = 1, \sum_{k=1}^{K} \lambda_k = 1.$$

Therefore, the linear multiplication of $(\varepsilon^d)^{\mathrm{T}}$, T^d, θ and λ maintains the condition of $\sum_{k=1}^{K} \vartheta_k^d = 1$ without normalization of ϑ^d. With ϑ^d, the topic proportions of the document d, the remaining part of the generative process is just familiar with LDA [5].

C. Compared with LDA and other Topic Models

As shown above, in TWDA, we introduce a novel way to model the semi-structured documents by leveraging the text data and the structured information (observed tags) in the documents. In LDA, the topic distribution is drawn from a hyperparameter, without considering the given tags. However, the tag information is useful for the generation of topic proportions. The proposed model in this paper can easily degenerate into LDA when we ignore the tags in a document. In this case, T^d will be a binary row vector whose last dimension equals to 1 and the others are 0, and $(\frac{\theta}{\lambda})$ is simplified to λ. Thus,

$$\begin{aligned} \vartheta^d &= (\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}) \\ &= \lambda. \end{aligned}$$

The topic distribution of d is simplified to λ, and as we shown above, λ is draw by a Dirichlet prior μ. It means that the topic proportions for the document d as a draw from a Dirichlet distribution which is the basic assumption of LDA [5].

It is important to distinguish TWDA from the Author-Topic Model [26]. In the author-topic model, the words w is chose only by one of the given tags' distribution, while in TWDA, for word w, all the observed tags in the document would make the contributions.

TWDA has the capability to deal with not only the multi-tag corpus but also non-tag corpus. For non-tag corpora, with the Dirichlet prior μ, TWDA may degenerate into LDA. For multi-tag corpus, there are two main differences between PLDA and TWDA. First, the generative process in PLDA assumes that a word in a document d is generated only by choosing one of the observed tags, while TWDA assumes the topic distribution of d obtained by weighted average of all the observed tags' topic distributions and a Dirichlet prior λ, which means that each word in d can be affected by some of the tags, which is more reasonable since the tags can be relevant. Second, a global topic distribution over tags (θ in Figure 1) can be obtained in TWDA. The benefit of this feature is that we can analyze the characteristics of the tag set of the corpus.

D. Learning and Inference

In this model, we treat π, μ, η, θ and ψ as unknown constants to be estimated, and use a variational expectation-maximization (EM) procedure to carry out approximate maximum likelihood estimation.

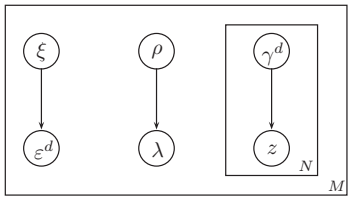


Fig. 2. Graphical model representation of the variational distribution used to approximate the posterior in TWDA

1) Variational E-step: Given the document d, we can easily get the posterior distribution of the latent variables in the proposed model, as:

$$p(\varepsilon^d, \mathbf{z}|\mathbf{w}^d, T^d, \theta, \eta, \psi, \pi, \mu) = \frac{p(\varepsilon^d, \mathbf{z}, \mathbf{w}^d, T^d|\theta, \eta, \psi, \pi, \mu)}{p(\mathbf{w}^d, T^d|\theta, \eta, \psi, \pi, \mu)}. \tag{1}$$

In Eq. (1), integrating over ε and summing out z, we easily obtain the marginal distribution of d:

$$\begin{aligned} &p(\mathbf{w}^d, T^d|\eta, \theta, \psi, \pi, \mu) = p(\mathbf{t}^d|\eta) \int p\left(\varepsilon^d|(T^d \times \pi)\right) \cdot \\ &p(\lambda|\mu) \prod_{i=1}^{N} \sum_{z_i^d=1}^{K} p(z_i^d|(\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda})) \cdot \\ &p(w_i^d|z_i^d, \psi_{1:K})\ d\varepsilon^d. \end{aligned}$$

As with LDA[24], it is not efficiently computable. Thus, we make use of mean-field variational EM algorithm [1] to efficiently obtain an approximation of this posterior distribution of the latent variables in TWDA. We maximize the evidence lower bound(ELBO) $\mathcal{L}(\cdot)$ [4] using Jensen's inequality, and for a document d we have the form:

$$\begin{aligned} &\mathcal{L}(\xi_{1:l^d}, \gamma_{1:K}, \rho_{1:K}; \eta_{1:L}, \pi_{1:L}, \mu_{1:K}, \theta_{1:L}, \psi_{1:K}) \\ &= E[\log p(T_{1:l^d}|\eta_{1:L})] + E[\log p(\varepsilon^d|T^d \times \pi)] \\ &+ E[\log p(\lambda^d|\mu)] + \sum_{i=1}^{N} E[\log p(z_i|(\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}))] \\ &+ \sum_{i=1}^{N} E[\log p(w_i|z_i, \psi_{1:K})] + H(q), \end{aligned}$$

where ξ is a l^d-dimensional Dirichlet parameter vector, ρ is a $1 \times K$ vector and γ is $1 \times K$ vector, all of which are variational parameters of variational distribution shown in Figure 2, and $H(q)$ indicates the entropy of the variational distribution:

$$H(q) = -E[\log q(\varepsilon^d)] - E[\log q(\lambda)] - E[\log q(z)].$$

Here the exception is taken with respect to a variational distribution $q(\varepsilon^d, q(\lambda^d), z_{1:N})$, and we choose the following fully factorized distribution:

$$\begin{aligned} &q(\varepsilon^d, \lambda^d, z_{1:N}|\xi_{1:L}, \rho_{1:K}, \gamma_{1:K}) \\ &= q(\varepsilon^d|\xi) q(\lambda^d|\rho) \prod_{i=1}^{N} q(z_i|\gamma_i). \end{aligned}$$

However, the term of the expected log probability of a topic assignment

$$E[\log p(z_i|(\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}))]$$
$$= \sum_{k=1}^{K} \gamma_{ik} E[\log((\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}))_k]$$

could be difficult to compute, because of tag-weighted topic assignment which is used in TWDA. Thus we use Jensen's inequality:

$$E[\log((\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}))_k]$$
$$= E[\log(\sum_{i=1}^{l^d-1} \varepsilon_i^d \theta_k^{(i)} + \varepsilon_{l^d}^d \lambda_k)]$$
$$\geq E[\sum_{i=1}^{l^d-1} \varepsilon_i^d \log \theta_k^{(i)} + \varepsilon_{l^d}^d \cdot \log \lambda_k]$$
$$= \sum_{i=1}^{l^d-1} \log \theta_k^{(i)} E[\varepsilon_i^d] + E[\varepsilon_{l^d}^d \cdot \log \lambda_k],$$

where the expression of $\theta^{(i)}$, $i \in \{1, \cdots, l^d - 1\}$, means the i-th tag's topic assignment vector, corresponding to the i-th row of Θ^d.

The first expectation is $E[\varepsilon_i^d] = \xi_i / \sum_{j=1}^{l^d} \xi_j$, and because the variational distribution is fully factorized, so the second expectation is

$$E[\varepsilon_{l^d}^d \cdot \log \lambda_k] = E[\varepsilon_{l^d}^d] \cdot E[\log \lambda_k],$$

where $E[\varepsilon_{l^d}^d] = \xi_{l^d} / \sum_{j=1}^{l^d} \xi_j$, and $E[\log \lambda_k] = \Psi(\rho_k) - \Psi(\sum_{j'=1}^{K} \rho_{j'})$. Thus, for the document d,

$$\sum_{i=1}^{N} E[\log p(z_i|(\varepsilon^d)^{\mathrm{T}} \times T^d \times (\frac{\theta}{\lambda}))] = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \cdot$$
$$[\sum_{j=1}^{l^d-1} \log \theta_k^{(j)} \frac{\xi_j}{\sum_{j'=1}^{l^d} \xi_{j'}} + (\Psi(\rho_k) - \Psi(\sum_{j'=1}^{K} \rho_{j'})) \frac{\xi_{l^d}}{\sum_{j=1}^{l^d} \xi_j}].$$

For a single document d, the variational parameters include ξ^d, ρ^d and γ_{ik}. First, we maximize $\mathcal{L}(\cdot)$ with respect to the variational parameters to obtain an estimate of the posterior.

Optimization with respect to ξ: We first maximize $\mathcal{L}(\cdot)$ with respect to ξ_i for the document d. Maximize the terms which contain ξ:

$$\begin{aligned}\mathcal{L}_{[\xi]} &= \sum_{i=1}^{l^d} (\sum_{l'=1}^{L+1} \pi_{l'} T_{il'}^d - 1)(\Psi(\xi_i) - \Psi(\sum_{j'=1}^{l^d} \xi_{j'})) \\ &+ \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \cdot \sum_{j=1}^{l^d} C_k^{(j)} \xi_j / \sum_{j'=1}^{l^d} \xi_{j'} \\ &- \log \Gamma(\sum_{i=1}^{l^d} \xi_i) + \sum_{i=1}^{l^d} \log \Gamma(\xi_i) \\ &- \sum_{i=1}^{l^d} (\xi_i - 1)(\Psi(\xi_i) - \Psi(\sum_{j'=1}^{l^d} \xi_{j'})), \end{aligned} \tag{2}$$

where

$$C_k^{(j)} = \begin{cases} \log \theta_k^{(j)} & j \in \{1, \cdots, l^d - 1\} \\ \Psi(\rho_k) - \Psi(\sum_{j'=1}^{K} \rho_{j'}) & j = l^d \end{cases}, \tag{3}$$

and $\Psi(\cdot)$ denotes the digamma function, the first derivative of the log of the Gamma function. The derivative of Eq. (2) with respect to ξ_i is

$$\begin{aligned}\mathcal{L}'(\xi_i) &= \Psi'(\xi_i) \left(\sum_{l=1}^{L+1} \pi_l T_{il}^d - \xi_i \right) - \Psi' \left(\sum_{j=1}^{l^d} \xi_j \right) \cdot \\ &\sum_{i=1}^{l^d} \left(\sum_{l=1}^{L+1} \pi_l T_{il}^d - \xi_i \right) + \sum_{i'=1}^{N} \sum_{k=1}^{K} \gamma_{i'k}^d \cdot \\ &(\frac{C_k^{(j)} (\sum_{j=1}^{l^d} \xi_j) - \sum_{j=1}^{l^d} C_k^{(j)} \xi_j}{(\sum_{j'=1}^{l^d} \xi_{j'})^2}). \end{aligned} \tag{4}$$

Here we use gradient descent method to find the ξ to make the maximization of $\mathcal{L}_{[\xi]}$.

Optimization with respect to ρ: Next, we maximize $\mathcal{L}(\cdot)$ with respect to ρ. The terms that involve the variational Dirichlet ρ are:

$$\begin{aligned}\mathcal{L}_{[\rho]} &= \sum_{i=1}^{K} (\mu_i - 1)(\Psi(\rho_i) - \Psi(\sum_{j=1}^{K} \rho_j)) - \log \Gamma(\sum_{j=1}^{K} \rho_j) \\ &+ \sum_{i=1}^{K} \log \Gamma(\rho_i) - \sum_{i=1}^{K} (\rho_i - 1)(\Psi(\rho_i) - \Psi(\sum_{j=1}^{K} \rho_j)) \\ &+ \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma_{ik} \cdot \frac{\xi_{l^d}}{\sum_{j=1}^{l^d} \xi_j} \cdot (\Psi(\rho_k) - \Psi(\sum_{j=1}^{K} \rho_j)). \end{aligned}$$

This simplifies to:

$$\begin{aligned}\mathcal{L}_{[\rho]} &= \sum_{i=1}^{K} (\Psi(\rho_i) - \Psi(\sum_{j=1}^{K} \rho_j)) \cdot \\ &(\mu_i - \rho_i + \sum_{n=1}^{N} \gamma_{ni} \cdot \frac{\xi_{l^d}}{\sum_{j=1}^{l^d} \xi_j}) \\ &- \log \Gamma(\sum_{j=1}^{K} \rho_j) + \sum_{i=1}^{K} \log \Gamma(\rho_i). \end{aligned}$$

Taking the derivative with respect to ρ_i and setting it to zero, we obtain a maximum at:

$$\rho_i = \mu_i + \sum_{n=1}^{N} \gamma_{ni} \cdot \frac{\xi_{l^d}}{\sum_{j=1}^{l^d} \xi_j}. \tag{5}$$

Optimization with respect to γ: Adding the Lagrange multipliers to the terms which contain γ_{ik}, taking the derivative with respect to γ_{ik}, and setting the derivative to zero yields, we obtain the update equation of γ_{ik}:

$$\gamma_{ik} \propto \psi_{k,v^{w_i}} \exp\{\sum_{j=1}^{l^d} C_k^{(j)} \frac{\xi_j}{\sum_{j'=1}^{l^d} \xi_{j'}}\}, \tag{6}$$

where v^{w_i} denotes the index of w_i in the dictionary. In E-step, we update the ξ, ρ and γ for each document with the initialized model parameters.

2) M-step: The M-step needs to update five parameters: η, the tagging prior probability, π, the Dirichlet prior of the tags' weights, θ, the topic distribution over all tags in the corpus, ψ, the probability of a word under a topic, and μ, a Dirichlet prior of model. Because each document's tag-set is observed, the Bernoulli prior η is unused included for model completeness. For a given corpus, the η_i is estimated by adding up the number of i-th tag which appears in the corpus.

For the document d, the terms that involve the Dirichlet prior π:

$$\begin{aligned}\mathcal{L}_{[\pi]} &= \log\Gamma(\sum_{i=1}^{l^d}(T^d\times\pi)_i) - \sum_{i=1}^{l^d}\log\Gamma\left((T^d\times\pi)_i\right) \\ &+ \sum_{i=1}^{l^d}\left((T^d\times\pi)_i - 1\right)(\Psi(\xi_i) - \Psi(\sum_{j=1}^{l^d}\xi_j)), \end{aligned} \quad (7)$$

where $(T^d\times\pi)_i = \sum_{l=1}^{L+1}\pi_l T^d_{il}$. We use gradient descent method by taking derivative of Eq. (7) with respect to π_l on the corpus to find the estimation of π. Taking derivatives with respect to π_l on the corpus, we obtain:

$$\begin{aligned}\mathcal{L}'_{[\pi_l]} = &\sum_{d=1}^{D}\Psi(\sum_{i=1}^{l^d}\sum_{l'=1}^{L+1}\pi_{l'}\cdot T^d_{il'})\cdot\sum_{i=1}^{l^d}T^d_{il} \\ &- \sum_{d=1}^{D}\sum_{i=1}^{l^d}\Psi(\sum_{l'=1}^{L+1}\pi_{l'}\cdot T^d_{il'})\cdot T^d_{il} + \sum_{d=1}^{D}\sum_{i=1}^{l^d}(\Psi(\xi_i) \\ &-\Psi(\sum_{j=1}^{l^d}\xi_j))\cdot T^d_{il}. \end{aligned} \quad (8)$$

The only term that involves θ is:

$$\mathcal{L}_{[\theta]} = \sum_{d=1}^{D}\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}\sum_{j=1}^{l^d}\log\theta_k^{(j)}\xi_j / \sum_{j'=1}^{l^d}\xi_{j'}, \quad (9)$$

where $\xi_j,\ j\in\{1,\cdots,l^d\}$ in the document d needs to be extended to $t^d_l\cdot\xi^d_l$, $l\in\{1,\cdots,L+1\}$ for convenient to simplify $\mathcal{L}_{[\theta]}$. With the Lagrangian of the Eq. (9), which incorporate the constraint that the K-components of θ_l sum to one, we obtain the estimation of θ over the whole corpus,

$$\theta_{lk}\propto\sum_{d=1}^{D}\sum_{i=1}^{N}\gamma^d_{ik}\frac{\xi^d_l t^d_l}{\sum_{l=1}^{L+1}(\xi^d_l t^d_l)}. \quad (10)$$

To maximize with respect to ψ, we isolate corresponding terms and add Lagrange multipliers:

$$\begin{aligned}\mathcal{L}_{[\psi]} &= \sum_{d=1}^{D}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{V}\gamma_{ik}(w^d)^j_i\log\psi_{kj} \\ &+ \sum_{k=1}^{K}\lambda_k(\sum_{j=1}^{v}\psi_{kj} - 1).\end{aligned}$$

Take the derivative with respect to ψ_{kj}, and set it to zero, we get:

$$\psi_{kj}\propto\sum_{d=1}^{D}\sum_{i=1}^{N}\gamma^d_{ik}(w^d)^j_i. \quad (11)$$

For the variational Dirichlet parameters μ, the involved terms are:

$$\begin{aligned}\mathcal{L}_{[\mu]} &= \sum_{d=1}^{D}(\log\Gamma(\sum_{j=1}^{K}\mu_j) - \sum_{i=1}^{K}\log\Gamma(\mu_i) \\ &+ \sum_{i=1}^{K}(\mu_i - 1)(\Psi(\rho^d_i) - \Psi(\sum_{j=1}^{K}\rho^d_j))). \end{aligned} \quad (12)$$

We can invoke the linear-time Newton-Raphson algorithm to estimate μ as same as the Dirichlet parameter described in LDA [5].

We summarize the variational expectation-maximization (EM) procedure of TWDA in Algorithm 1.

Algorithm 1 The variational expectation-maximization (EM) algorithm of TWDA

```
Input: a semi-structured corpora including totally V unique words, L unique tags, and the expected number K of topics
Output: Topic-word distributions ψ, Tag-topic distributions θ, π, μ, topic distribution ϑ^d and weight vector ε^d of each training document.
initialize π and μ.
initialize θ and ψ with the constraint of Σ_{k=1}^{K} θ_lk equals 1 and Σ_{i=1}^{V} ψ_ki equals 1.
repeat
    for for each document d do
        update ξ^d with Eqs. (2) and (4) using gradient descent method,
        update ρ with Eq. (5),
        update γ_ik with Eq. (6).
    end for
    update π with Eqs. (7) and (8) using gradient descent method,
    update μ with Eq. (12) by Newton-Raphson algorithm,
    update θ by Eq. (10),
    update ψ by Eq. (11).
until convergence
```

IV. Experimental Analysis

A. Experiment Settings

In the experiments of this work, we used two semi-structured corpora. The first one consists of technical papers of the Digital Bibliography and Library Project (DBLP) data set[3], which is a collection of bibliographic information on major computer science journals and proceedings. In this paper, we use a subset of DBLP that contains abstracts of D=27,435 papers, with W=70,062 words in the vocabulary and L=6,256 unique tags. The tags we used in DBLP include authors and keywords. And the second document collection is the data from Internet Movie Database (IMDB)[4]. The data set includes 12,091 movie storylines, 52,274 words after removing stop words, and 3,654 tags. These movies belong to 29 genres. And the tags we used contain directors, stars, time, and movie keywords.

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

[4]http://www.imdb.com

B. Results on Documents Modeling

In order to evaluate the generalization capability of the model, we use the perplexity score that described in [5]. There are two parts of the experiments.

First, We trained three latent variable models including LDA [5], CTM [3] and our TWDA, on the corpora of a set of movie documents in IMDB, to compare the generalization performance of the three models. In this part, LDA and CTM trains text data without taking advantage of tag information. We removed the stop words and conducted experiments using 5-fold cross-validation. Figure 3(a) demonstrates the perplexity results on the IMDB data set. Clearly, TWDA excels both CTM and LDA significantly and consistently, and the results show that TWDA works very well in semi-structured document modeling.

Second, in order to compare the performance of TWDA with other topic models which take advantage of the tag information, we trained TWDA, DMR[5], PLDA[6], Author Topic Model (ATM) [26], CTM, and LDA on the set of movie documents in IMDB and computed the perplexity on test data set as described in DMR [21]. Since CTM and LDA could not handle corpus with tags easily, in this experiment, we treated the given tags as word features for them. Figure 3(b) demonstrates the perplexity results of the five models on the IMDB data. The experiment results shows that TWDA is better than other models, and when T increases, DMR, CTM and LDA are running into over-fitting, while the trend of TWDA keeps going down and the perplexity is significantly lower than those of the baselines.

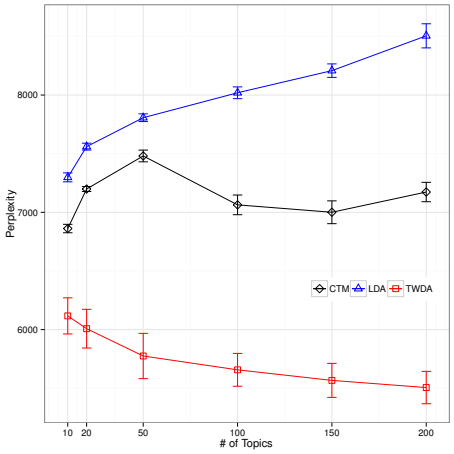
As PLDA [25] assumes that one of tags may optionally denote as a tag "latent" present on every document d, thus, we trained PLDA and TWDA over 1021 and 2041 topics on IMDB data set with 1020 tags, since in PLDA, each latent topic takes part in exactly one tag in a collection. As shown in [25], PLDA builds on Labeled LDA [24], and when it set one latent topic and one topic for each tag, it is approximately equivalent to Labeled LDA. For this case, we trained PLDA over 1021 topics. Figure 3(c) shows the perplexity results of TWDA and PLDA. As the results of Figure 3 shown, TWDA works very well compared with other topic models which make use of tag information.

C. Results on Tags prediction

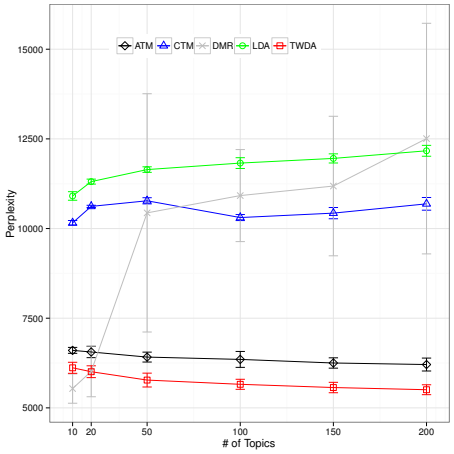
In the preceding section we demonstrated the performance of TWDA on the tags prediction by process the paper collection in DBLP. In addition to predicting the tags given a document, we evaluate the ability of the proposed model, compared with the Author Topic Model (ATM) [26] and DMR [21], to predict the tags of the document conditioned on words in the document. In this part, we treat the authors of each paper as the tags, and the abstract as the word features, and we predict the authors of one paper by modeling the paper abstract document data using ATM, DMR, and TWDA. For each model, we can evaluate the likelihood of the authors given the word features of one document, and rank each possible

[5]We used the Mallet code (http://mallet.cs.umass.edu/).

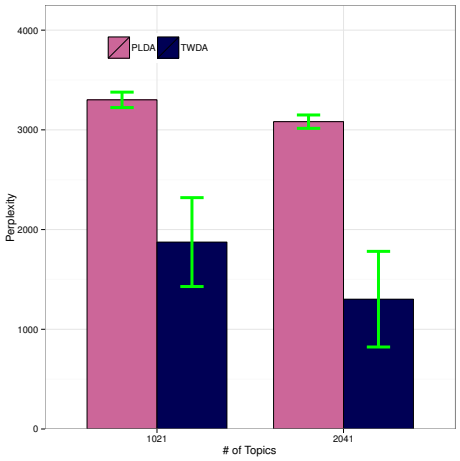
[6]We used the code in Stanford Topic Modeling Toolbox (http://www-nlp.stanford.edu/software/tmt/tmt-0.4/).



(a) Perplexity results for TWDA, LDA and CTM

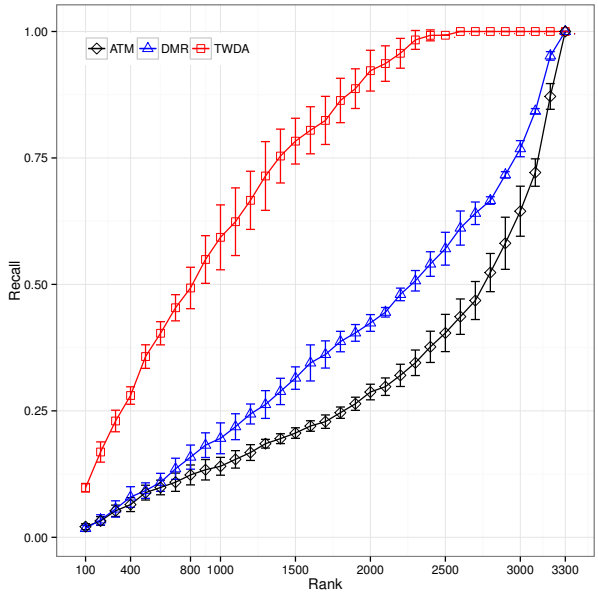


(b) Perplexity results for TWDA, DMR, ATM, LDA and CTM

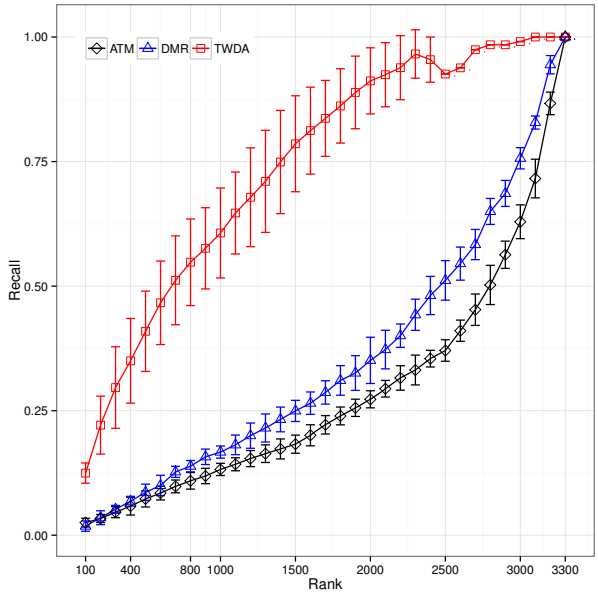


(c) Perplexity results for TWDA and PLDA

Fig. 3. Perplexity results of different models on IMDB corpora. LDA and CTM only use the words when training in (a), and add the tags as the word feature during the training process in (b).



(a) Results of Author retrieval recall for TWDA, DMR and ATM by setting K=100



(b) Results of Author retrieval recall TWDA, DMR and ATM by setting K=200

Fig. 4. Prediction results of TWDA, DMR and ATM for authors on DBLP corpora. We set the number of topic in the corpora to be 100 in (a) and 200 in (b).

author by the likelihood function of the author. First, for each model, we can get the topic distribution over a test document d_{test} given one author a. Then, we evaluate the $p(d_{test}|a)$ for d_{test} over each author a in the tags(authors) set by

$$p(d_{test}|a) = \prod_{i}^{N}(\sum_{z} p(z|a)p(w_i|z)).$$

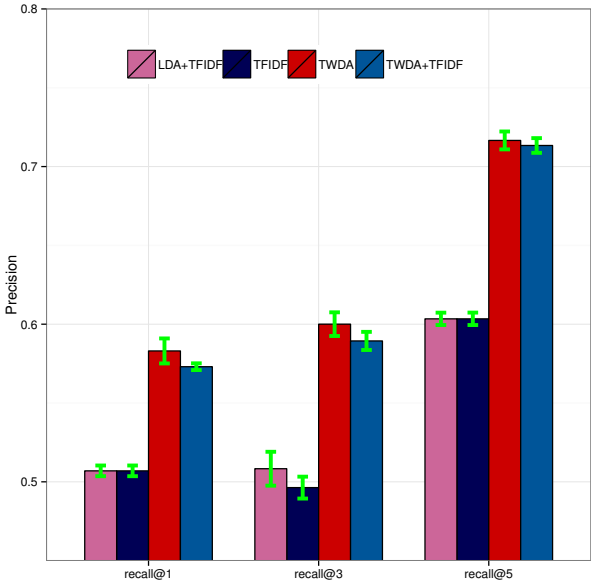


Fig. 5. Classification results of different features on recall@1, recall@3 and recall@5 with 5-fold cross-validation.

For DMR and ATM, the method which define $p(d_{test}|a)$ is shown as [21]. Note that the likelihoods for a given author over a document are not necessarily comparable among the topic models, as described in [21], however, what we are interested in is the ranking as same as [21].

We trained the three models on DBLP data set using 5-fold cross-validation and shows the recall when the topic in the corpora is set to be 100 and 200. Results are shown in Figure 4(a) and Figure 4(b). TWDA ranks authors consistently higher than ATM and DMR.

D. Results on Feature Construction for Classification

The next experiment is to test the classification performance utilizing feature sets generated by TWDA and other baselines. For the base classifier, we use LIBSVM [9] with Gaussian kernel and the default parameters. For the purpose of comparison, we trained four SVMs on tf-idf word features, features induced by a 30-topic LDA model and tf-idf word features, features generated by a TWDA model with the same number of topics, and features induced by a 30-topic TWDA model and tf-idf word features respectively.

In these experiments, we conducted multi-class classification experiments using the IMDB data set, which contains 29 genres. We calculated the evaluation metrics recall@1, recall@3 and recall@5 with the provided class tags of movies' genres, using 5-fold cross-validation. We report the movie classification performance of the different methods in Figure 5, where we see that there is significant improvement in classification performance when using LDA and TWDA comparing with only using tf-idf features, and TWDA outperforms both LDA and tf-idf in terms of recall@1, recall@3 and recall@5.

In order to show the classification performance better, we

TABLE I. CLASSIFICATION RESULTS OF DIFFERENT FEATURES ON F1-SCORE

F1-score	@1	@3	@5
TFIDF	0.5	0.41	0.39
LDA+TFIDF	0.5	0.42	0.39
TWDA	0.57	0.5	0.47
TWDA+TFIDF	0.58	0.5	0.47

TABLE II. SOME EXAMPLES OF THE NORMALIZED WEIGHTS AMONG THE ORIGINAL TAGS AND NOISE TAGS. THE TAGS IN RED ARE NOISE TAGS, AND THE NUMBERS ARE THE WEIGHT VALUES.

"Bug isolation via remote program sampling [19]"	
Ben Liblit: 0.185	Alex Aiken: 0.2257
aAlice X. Zheng: 0.228	Michael I. Jordan: 0.349
K. G. Shin: 0.01	

"Web question answering: is more always better? [13]"	
Susan Dumais: 0.986	Michele Banko: 0.0032
Eric Brill: 0.0038	Jimmy Lin: 0.0038
Andrew Ng: 0.0024	
R. Katz: 0.00018	

"Contextual search and name disambiguation in email using graphs [22]"	
Einat Minkov: 0.425	William W. Cohen: 0.342
Andrew Y. Ng: 0.128	
J. Ma: 0.033	*D. Ferguson: 0.07*

"A Sparse Sampling Algorithm for Near-Optimal Planning in Large Markov Decision Processes [16]"	
Michael Kearns: 0.296	Yishay Mansour:0.166
Andrew Y. Ng: 0.31	
J. Blythe: 0.089	*B. Adida: 0.027*
P. J. Modi: 0.1	

"The nested Chinese restaurant process and bayesian nonparametric inference of topic [2]"	
David M. Blei:0.46	Thomas L. Griffiths:0.186
Michael I. Jordan:0.225	
B. Clifford:0.031	*R. Szeliski:0.048*
X. Wang:0.05	

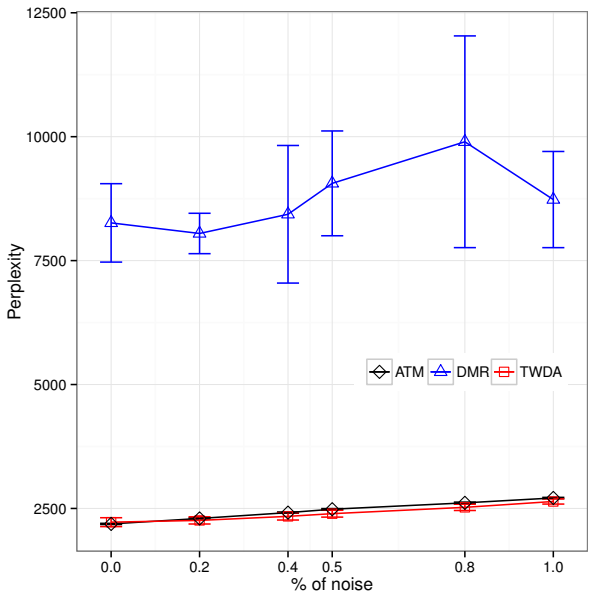
also calculated the evaluation metrics F-Measure (F1-score) by

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

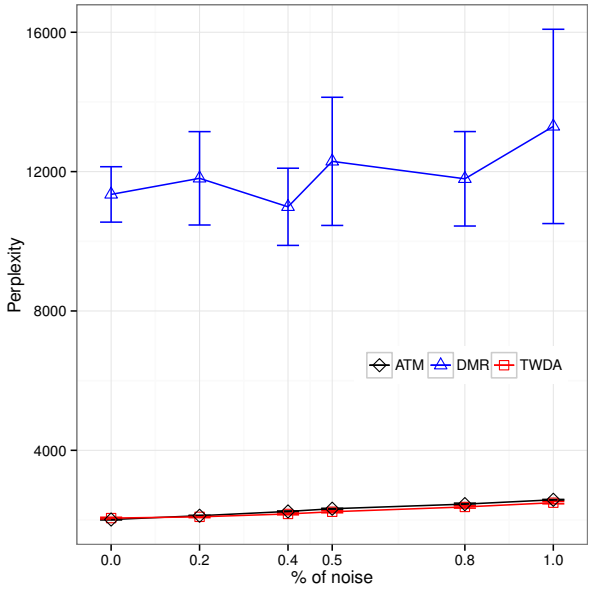
The results of F-Measure is reported in Table I. TWDA provides substantially better performance on F-Measure.

E. Results on Model Robustness

We demonstrated the performance of the proposed model on model robustness in the last part of experimental analysis. In this part, we measured and compared the perplexity when we added noise tags information to the test documents using DBLP data set. Respectively, we randomly added 20%, 40%, 50%, 80% and 100% noise tags into a test document and then calculated the perplexity. For example, if a paper document in DBLP has five authors, adding 20% noise is that we randomly selected one author (not appeared in the paper) from the author set of the DBLP corpora and added into the paper as a noise author.



(a) K=100



(b) K=200

Fig. 6. The Results of adding noise to different models(ATM, DMR and TWDA). (a) set K=100, and (b) set K=200. Steady trending means a good performance on model robustness.

And in some real applications, the noise tags appeared in a document may have some relevance to the real tags. So in this experiment, we selected the noise tags from the author-tag set to meet the real applications to some extent. In this experiment, the DBLP corpora contains more than 6,000 tags, the noise tags we added into a test document would be very

sparse for the whole tag set in the corpora. So, we added the different percentages noise tags into the test document to show the trend of perplexity as the noise content increases. Figure 6 shows that both TWDA and ATM have a more steady trend as the noise level increases, compared with DMR.

Table II shows some examples about the weights between the original tags and noise tags. The red tags are the noise added into the test data, and the values behind are the weights among the tags we inference from the TWDA model. Note that, we showed the weight values after normalized. As the results shown, TWDA has a good performance on model robustness, for the weight values of the noise tags are much smaller than the other original tags. In some applications, we can use the proposed model to rank the tags given in a document, which would be a good approach to tag recommendation and annotation.

V. Conclusion

In this paper, we propose a topic model called tag-weighted Dirichlet allocation, which builds on LDA and provides a probabilistic approach for mining semi-structured documents. This model shows a novel method to generate and analyze the topic proportions of a document using normalized weights of a Dirichlet prior and the structure information (observed tags) in the document. And as shown in the experiments, it not only provides significant improvement in term of document modeling compared to other topic models, but also gives a way to predict the latent tags for tag-unknown documents. Meanwhile, the TWDA can handle the multi-tag documents and non-tag documents at the same time, since when it comes to non-tag documents, TWDA degenerates into LDA easily. In the future, more work should to be done on the effective and efficient solutions of large scale semi-structured documents and the different practical areas (e.g., image classification and annotation, video retrieval).

Acknowledgment

We thank the anonymous reviewers for helpful comments. This work was supported by National Science Foundation of China (61003140 and 61033010).

References

[1] Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*. *J. Electronic Imaging*, 16(4):049901, 2007.

[2] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2):7:1–7:30, February 2010.

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

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

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

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

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

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

[9] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[10] Jonathan Chang and David M. Blei. Relational topic models for document networks. *Journal of Machine Learning Research - Proceedings Track*, 5:81–88, 2009.

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

[12] 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.

[13] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. Web question answering: is more always better? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 291–298, New York, NY, USA, 2002. ACM.

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

[15] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling social annotation data with content relevance using a topic model. In *NIPS*, pages 835–843, 2009.

[16] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Mach. Learn.*, 49(2-3):193–208, November 2002.

[17] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, pages 897–904, 2008.

[18] Shuangyin Li, Jiefei Li, and Rong Pan. Tag-weighted topic model for mining semi-structured documents. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[19] Ben Liblit, Alex Aiken, Alice X. Zheng, and Michael I. Jordan. Bug isolation via remote program sampling. *SIGPLAN Not.*, 38(5):141–154, May 2003.

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

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

[22] Einat Minkov, William W. Cohen, and Andrew Y. Ng. Contextual search and name disambiguation in email using graphs. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 27–34, New York, NY, USA, 2006. ACM.

[23] James Petterson, Alexander J. Smola, Tibério S. Caetano, Wray L. Buntine, and Shravan Narayanamurthy. Word features for latent dirichlet allocation. In *NIPS*, pages 1921–1929, 2010.

[24] 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.

[25] Daniel Ramage, Christopher D. Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[26] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1):4:1–4:38, January 2010.

[27] Markus Tresch, Neal Palmer, and Allen Luniewski. Type classification of semi-structured documents. In *VLDB*, pages 263–274, 1995.

[28] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA, 2006. ACM.

[29] Jeonghee Yi and Neel Sundaresan. A classifier for semi-structured documents. In *KDD*, pages 340–344, 2000.

[30] Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *ICML*, page 158, 2009.