Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Tag-Weighted Topic Model for Mining Semi-Structured Documents

Shuangyin Li, Jiefei Li and Rong Pan*
Department of Computer Science
Sun Yat-sen University, Guangzhou, China
{lishyin@mail2., lijiefei@mail2., panr@}sysu.edu.cn

Abstract

In the last decade, latent Dirichlet allocation (LDA) successfully discovers the statistical distribution of the topics over a unstructured text corpus. Meanwhile, more and more document data come up with rich human-provided tag information during the evolution of the Internet, which called semi-structured data. The semi-structured data contain both unstructured data (e.g., plain text) and metadata, such as papers with authors and web pages with tags. In general, different tags in a document play different roles with their own weights. To model such semi-structured documents is nontrivial. In this paper, we propose a novel method to model tagged documents by a topic model, called Tag-Weighted Topic Model (TWTM). TWTM is a framework that leverages the tags in each document to infer the topic components for the documents. This allows not only to learn document-topic distributions, but also to infer the tag-topic distributions for text mining (e.g., classification, clustering, and recommendations). Moreover, TWTM automatically infers the probabilistic weights of tags for each document. We present an efficient variational inference method with an EM algorithm for estimating the model parameters. The experimental results show that our TWTM approach outperforms the baseline algorithms over three corpora in document modeling and text classification.

1 Introduction

Large collections of text with kinds of metadata appears in many Web applications. The 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 area, such as information retrieval, artificial intelligence and machine learning. The tags can be more important than the text data in document mining. For example, IMDB, the world's most popular and authoritative source for movie, TV and celebrity content. Each movie has lots of tags, e.g., director, writers, stars, country, language and so on, and a storyline as text data. Given a movie with a tag "Dick Martin", we may have an opinion that it has a higher probability to be a comedy, without read all its storyline or watch it.

In past decade, topic models have been used to be a powerful method of management of document corpus, such as latent Dirichlet allocation (LDA) [Blei *et al.*, 2003], which assigns topics to corpora and generates topic distributions over words given corpus. 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. Researchers have proposed approaches to deal with tags or labels [Mimno and McCallum, 2008; Ramage *et al.*, 2009; 2011]. However, there are still two main problems. Firstly, in a document, the importance of the tags can be different in terms of document modeling. For instance, in the domain of scientific paper modeling, there are three authors in a specific paper, which can be three tags of the paper. In most cases, the first author would make more contribution than the third author. In fact, different authors have different background. Therefore the weights of the authors can affect the formation of the topic distribution of the document. Secondly, tags in a corpus are diverse and evolving. For a news story in NYTimes, the tags set contains various types, such as reporters and category, and both have an impact on topics assignment with different weights. It is clear that how to take advantage of all the available tags is a challenge when sundry tags are involved.

In this work, we focus on document collections where documents are tagged. We propose a tag-weighted topic model (TWTM) to represent the text data and the various tags with weights to evaluate the importance of the tags. Besides, TWTM also infers the topic distributions of tags. The weights of observed tags in document, which we infer from data set, give us an opportunity to provide a method to rank the tags. Compared to LDA, TWTM can conduct document modeling with lower perplexity, and construct more discriminant features for classification.

The proposed TWTM has three principal contributions.

1. It is a novel topic modeling method to model the semi-structured data, not only generating the topic distributions over words, but also infering topic distributions of

*Corresponding author

tags.

2. TWTM automatically infers a topic distribution for each individual document directly, with a function of tag-weighted topic assignment.
3. Instead of constructing a new model for new tags in the corpus, TWTM provides a framework to model a variety of tags, which shows the extensibility of the model.

The rest of the paper is organized as follows. In Section 2, we first analyze and discuss related works. In Section 3, we first introduce basic notation and terminology, present a novel topic model, and give the method of learning and inference. In Section 4, we present the experimental results on three domains to show the performance of the proposed method in document modeling and text classification. We conclude and discuss further research directions in Section 5.

2 Related Works

There are many topic models proposed and shown to be useful on document analyzing, such as in [Petterson *et al.*, 2010; Hofmann, 1999; Blei *et al.*, 2003; Blei and McAuliffe, 2007; Boyd-Graber and Blei, 2010; Chang and Blei, 2009], which have been applied to many areas, including document clustering and classification [Cai *et al.*, 2008], and information retrieval [Wei and Croft, 2006]. They are extended to many other topic models for different situation of applications in analyzing text data [Iwata *et al.*, 2009; Lacoste-Julien *et al.*, 2008; Zhu *et al.*, 2009]. However, most of these models only consider the textual information while ignore human-provided tag information.

Several models have been proposed to take advantage of tags or labels, such as Labeled LDA [Ramage *et al.*, 2009], DMR [Mimno and McCallum, 2008] and PLDA [Ramage *et al.*, 2011], or modeling relationships among several variables, such as Author-Topic Model [Rosen-Zvi *et al.*, 2004]. Labeled LDA [Ramage *et al.*, 2009] 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. Thus, it is hard to infer the topic distribution of labels, that is very useful for text classification in some web applications. And in Author Topic Model, it obtains the topic distributions of authors, without gives the importance between the given authors in a document. DMR [Mimno and McCallum, 2008] is a Dirichlet-multinomial regression topic model that includes a log-linear prior on document-topic distributions that is a function of given features of the document. However, there is a lack of the information of the tag weights in DMR as well as Author-Topic Model [Rosen-Zvi *et al.*, 2004]. PLDA [Ramage *et al.*, 2011] provides another way of modeling the labels' text data, that assumes the generation of words of topics assignment is limited by only one of the given labels, and estimates an appropriate size for each per-label topic set automatically using Dirichlet process.TMBP [Deng *et al.*, 2011] and cFTM [Chen *et al.*, 2012] give methods to make use of the contextual information of documents to model the topic assignment. TMBP is a topic model with biased propagation to leveraging contextual information, the authors and venue. However, predefining the weights of the author and venue information on word assignment limits the usefulness in many real applications, and cFTM with the strict assumption of either author or venue be chosen for one word topic assignment lack of scalability when there are various types of contextual information.

3 TWTM Model and Algorithms

In this section, we will mathematically define the tag-weighted topic model (TWTM), and discuss the learning and inference methods. In this model, just like the latent Dirichlet allocation (LDA) [Blei *et al.*, 2003], we treat the words of document as arising form a set of latent topics.

3.1 TWTM

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 $l^d \times L$ matrix T^d, where l^d is the number of tags in document d. 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[1] is the j-th tag of the tag set in the corpus D. 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.

In the past topic models, each document d is typically characterized by a distribution over topics, θ^d , and each topic k is represented by ψ_k, over words in a vocabulary. Take LDA [Blei *et al.*, 2003] for an example, the generative process of topic distribution of document d is assumed as follows.

$$\text{Choose } \theta^d \sim \text{Dirichlet}(\alpha),$$
$$\text{and choose } z_{ni} \sim \text{Multinomial}(\theta^d),$$

where α is the hyperparameter. In LDA, the topic distribution θ^d is drawn from a hyperparameter, α without considering the tags' information. However, the human-provided tag information is useful to impact on the formation of θ^d.

In this paper, we use ϑ^d to denote the topic distribution of document d, controlled by observed tags, as shown in Figure 1. And, let θ represent a $L \times K$ topic distribution matrix over the tag set, where L is the size of the tag set in corpus D, and K is the number of topics. Let ψ represent a $K \times V$ distribution matrix over words in the dictionary, where V is the number of words in the dictionary of D.

Similar to LDA, TWTM models the document d as a mixture of underlying topics and generates each word from one topic. While the topic proportions ϑ^d of document d in TWTM is a mixture of tags' (given/specified) topic distributions, not restrict to be defined only over a hyperparameter in LDA. Meanwhile, the novel method we propose to model the

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

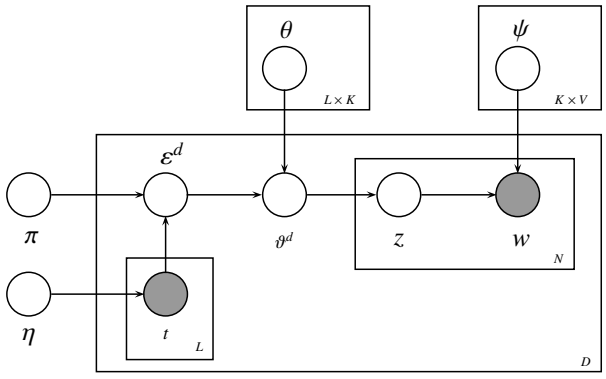

Figure 1: Graphical model representation for TWTM, where θ is distribution matrix of the whole tags, ψ is distribution matrix of words. ϑ^d indicates the topic components for each document.

topic proportions ϑ^d is different from neither cFTM nor Labeled LDA, or other topic models, such as PLDA and DMR.

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

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

In this process, Dir$(\cdot)$ designates a Dirichlet distribution, Mult$(\cdot)$ is a multinomial distribution, and π is a $L \times 1$ column vector, a Dirichlet prior.

Note that ε^d indicates the weight vector of the observed tags in constituting the topic proportions of document d. Furthermore, ε^d is drawn from a Dirichlet prior that is result of the matrix multiplication of $T^d \times \pi$. Clearly, the result of $T^d \times \pi$ will be a $(l^d \times 1)$ vector of which the dimension is depended on the number of observed tags in document d.

In Step 3, for one document d, we first generate the document's tags t_l^d using a Bernoulli coin toss with a prior probability η_l, as shown in (a). Then, after draw out the ε^d, we generate the ϑ^d through ε^d, T^d and θ, that will be discussed in section 2.3. The remaining part of the generative process is just familiar with the traditional LDA model [Blei *et al.*, 2003].

As shown above, TWTM assumes a novel way to model the topic proportions of semi-structured document by document-special tags and text data. The main key which discussed in this paper is the tag's weight topic assignment, which provides an effective and directly method to infer the weight of tags.

3.2 Tag-Weighted Topic Assignment

As we expect that all the observed tags in document d make contributions to infer the topic distribution ϑ^d of the document, and meanwhile, it is expected that different tags works corresponding to their weights. For example, a blog with tags of an author, blog's date, blog category and blog's url. Clearly, compared to other tags, the author plays the most important role in constituting the topic components of this blog.

The function of how to leverage the tag information or contextual for modeling topic distribution of document is defined as follows:

$$\vartheta \longleftarrow f(Tag_1, \cdots, Tag_l),$$

where $f(\cdot)$ is the way of making use of the tag information. Topic models with tag information or contextual in the past are take advantage of different $f(\cdot)$. In TWTM, we assume that ϑ^d is made up by all the observed tags with their own weight. Figure 1 shows that how TWTM works in probabilistic graphical model. As is shown in Figure 1, ϑ^d is controlled by two sides, the topic distributions over tags θ, and the weights of the given tags of document d.

It is important to distinguish TWTM from the Author-Topic Model [Rosen-Zvi *et al.*, 2004]. In the author-topic model, words w is chose by the distribution only from one of the given tags, while in TWTM, for word w, all the observed tags in the document would make the contributions.

The $f(\cdot)$ in the proposed model is assumed as this, for document d,

$$f(\vartheta^d) = (\varepsilon^d)^{\mathrm{T}} \times T^d \times \theta,$$

where the linear multiplication of $(\varepsilon^d)^{\mathrm{T}}$, T^d and θ maintains the condition of $\sum_k^K \vartheta_k^d = 1$ without normalization of ϑ^d, since that ε^d and θ satisfies

$$\sum_i^{l^d} \varepsilon_i^d = 1, \sum_k^K \theta_{lk} = 1.$$

Firstly, we pick out the topic distributions of the given tags in document d by $T^d \times \theta$, for T^d is a $l^d \times L$ matrix and θ is a $L \times K$ matrix. Here we define

$$\Theta^d = T^d \times \theta,$$

where the Θ^d is a $l^d \times K$ topic distribution matrix of given tags in d as sub-components of θ.

Secondly, ε^d is the weight vector of observed tags in d, and each dimension of ε^d represents the weight or importance associated to the corresponding tag. So, ϑ^d is mixed by Θ^d with corresponding weight values.

$$\vartheta^d = (\sum_i^{l^d} \varepsilon_i^d \Theta_{i1}^d, \ldots, \sum_i^{l^d} \varepsilon_i^d \Theta_{ij}^d, \ldots, \sum_i^{l^d} \varepsilon_i^d \Theta_{iK}^d).$$

With ϑ^d, TWTM generates all the words of document d with the assumption of bag-of-words, which is familiar with the LDA. It is worth to note that the generative process in PLDA [Ramage *et al.*, 2011] is just different from the TWTM model, where the assumption of topic assignment for d in PLDA is that the word generated by choosing only one of the observed tags' index, while TWTM assumes the topic distribution of d obtained by weighted average of all the observed tags' topic distributions.

3.3 Model Learning and Inference

In topic models, the key inferential problem that we need to solve is that of computing the posterior distribution of the hidden variables given a document d. Given a document d, we can easily get the posterior distribution of the latent variables in the proposed model, as:

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

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

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

However, this posterior distribution of the hidden variables ε and z is intractable to compute in general. A variety of algorithms have been used to estimate the parameters of topic models, such as Monte Carlo Markov chain(MCMC) sampling techniques [Andrieu *et al.*, 2003; Griffiths and Steyvers, 2004], variational methods [Attias, 1999] and others methods [Asuncion *et al.*, 2009; Sato and Nakagawa, 2012]. For sampling methods, actually, because of the nonconjugacy of the function of tag-weighted topic assignment, we have to appeal to a tailored solution of MCMC [Blei and Lafferty, 2005]. Such a technique will impede the requirement of convergence properties and speed, especially when the corpus comprise millions of words. Thus, we make use of mean-field variational EM algorithm [Bishop and Nasrabadi, 2007] to efficiently obtain an approximation of this posterior distribution of the latent variables in the TWTM model. In mean-field variational inference, we minimize the KL divergence between the variational posterior probability and the true posterior probability through by maximizing the evidence lower bound (ELBO) $\mathcal{L}(\cdot)$ [Blei and McAuliffe, 2007]. For a single document d, we obtain the $\mathcal{L}(\cdot)$ using Jensen's inequality:

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

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

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

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

$$q(\varepsilon^d, z_{1:N}|\xi_{1:L}, \gamma_{1:K}) = q(\varepsilon^d|\xi) \prod_{i=1}^{N} q(z_i|\gamma_i).$$

The dimension of parameter ξ is changed for different documents.

The way of tag-weighted topic assignment leads to difficulty in computing the expected log probability of a topic assignment:

$$\begin{aligned}&E[\log p(z_i|(\varepsilon^d)^{\mathrm{T}} \times T^d \times \theta)] \\ &= \textstyle\sum_k^K \gamma_{ik} E[\log((\varepsilon^d)^{\mathrm{T}} \times T^d \times \theta)_k].\end{aligned}$$

To preserve the lower bound on the log probability, we upper bound the log normalizer using Jensen's inequality again:

$$\begin{aligned}&E[\log((\varepsilon^d)^{\mathrm{T}} \times T^d \times \theta)_k] \\ =&E[\log \sum_i^{l^d} \varepsilon_i^d \theta_k^{(i)}] \\ \geq&E[\sum_i^{l^d} \varepsilon_i^d \log \theta_k^{(i)}] \\ =&\sum_i^{l^d} \log \theta_k^{(i)} E[\varepsilon_i^d],\end{aligned}$$

where the expression of $\theta^{(i)}$, $i \in \{1, \cdots, l^d\}$, means the i-th tag's topic assignment vector, corresponding to the i-th row of Θ^d. Note that the expectation of Dirichlet random variable is $E[\varepsilon_i^d] = \xi_j / \sum_j^{l^d} \xi_j$. Thus, for document d,

$$\begin{aligned}&\textstyle\sum_{i=1}^N E[\log p(z_i|(\varepsilon^d)^{\mathrm{T}} \times T^d \times \theta)] \\ &= \textstyle\sum_i^N \sum_k^K \gamma_{ik} \cdot \sum_j^{l^d} \log \theta_k^{(j)} \xi_j / \sum_{j'}^{l^d} \xi_{j'}.\end{aligned}$$

Then, we maximize the lower bound $\mathcal{L}(\xi, \gamma; \eta, \pi, \theta, \psi)$ with respect to the variational parameters ξ and γ, using a variational expectation-maximization(EM) procedure.

In particular, by computing the derivatives of the $\mathcal{L}(\cdot)$ and setting them equal to zero, we obtain the following update equations:

$$\gamma_{ik} \propto \psi_{k,v^{w_i}} \exp\{\textstyle\sum_{j=1}^{l^d} \log \theta_k^{(j)} \frac{\xi_j}{\sum_{j'=1}^{l^d} \xi_{j'}}\}, \quad (2)$$

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

$$\psi_{kj} \quad \propto \textstyle\sum_d^D \sum_i^N \gamma_{ik}^d (w^d)_i^j. \quad (3)$$

$$\theta_{lk} \quad \propto \textstyle\sum_d^D \sum_i^N \gamma_{ik}^d \frac{\xi_l^d t_l^d}{\sum_l^L (\xi_l^d t_l^d)}. \quad (4)$$

The ξ_j, $j \in \{1, \cdots, l^d\}$, in document d needs to be extended to $t_l^d \cdot \xi_l^d$, $l \in \{1, \cdots, L\}$ for convenient to simplify $\mathcal{L}_{[\theta]}$.

ξ_i can be implemented using gradient descent method, by taking the derivative of the terms:

$$\begin{aligned}\mathcal{L}_{[\xi]} = \quad &\textstyle\sum_i^{l^d} (\sum_{l'}^L \pi_{l'} T_{il'}^d - 1)(\Psi(\xi_i) - \Psi(\sum_{j'}^{l^d} \xi_{j'})) + \\ &\textstyle\sum_i^N \sum_k^K \gamma_{ik} \cdot \sum_j^{l^d} \log \theta_k^{(j)} \frac{\xi_j}{\sum_{j'}^{l^d} \xi_{j'}} - \\ &\log \Gamma(\textstyle\sum_i^{l^d} \xi_i) + \sum_i^{ld} \log \Gamma(\xi_i) - \\ &\textstyle\sum_i^{l^d} (\xi_i - 1)(\Psi(\xi_i) - \Psi(\sum_{j'}^{l^d} \xi_{j'})), \quad (5)\end{aligned}$$

where $\Psi(\cdot)$ denotes the digamma function, the first derivative of the log of the Gamma function.

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

$$\mathcal{L}_{[\pi]} = \log\Gamma\left(\sum_i^{l^d}(T^d\times\pi)_i\right) - \sum_i^{l^d}\log\Gamma\left((T^d\times\pi)_i\right) + \sum_i^{l^d}\left((T^d\times\pi)_i - 1\right)\left(\Psi(\xi_i) - \Psi(\sum_j^{l^d}\xi_j)\right), \quad (6)$$

where $(T^d\times\pi)_i = \sum_i^{l^d}\sum_l^L \pi_l T_{il}^d$. We can also use gradient descent method by taking derivatives of Eq. (6) with respect to π_l on the whole corpus to find the estimation of π. Because each document's tags-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 all documents.

We show the variational expectation-maximization(EM) procedure of TWTM as follows. At the beginning, TWTM initializes the parameter π ,θ and ψ, with the constraint of $\sum_k^K \theta_l$ equals 1 and $\sum_i^V \psi_k$ equals 1. In the E-step, for each document d, we update the variational parameter ξ and γ with Eqs. (5) and (2) to maximize the $\mathcal{L}(\xi,\gamma;\eta,\pi,\theta,\psi)$. In the M-step, π ,θ and ψ are updated as in Eqs. (6), (4), and (3). We repeatedly the update steps until the convergence condition of likelihood is satisfied or maximum number of iterations is reached.

4 Experimental Results

In the experiments of this work, we used three semi-structured corpora. The first one consists of technical papers of the Digital Bibliography and Library Project (DBLP) dataset[2], 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=8,212 papers, with W=34,967 words in the vocabulary and L=1,893 unique tags. The tags we used in DBLP include authors, time, and keywords. The second corpus is The New York Times[3] news stories in March 2011, which is made up of D=4,307 stories, in which we treat 31 days of March 2011 as tags. And the last document collection is the data from Internet Movie Database (IMDB)[4]. The dataset includes 12,091 movie storylines, 52,274 words after removing stop words, and 3,654 tags which contain directors, stars, time, and movie keywords. And these movies belong to 29 genres.

4.1 Document Modeling

In order to evaluate the generalization capability of the model, we use the perplexity score that described in [Blei *et al.*, 2003]. We trained three latent variable models including LDA [Blei *et al.*, 2003], CTM [Blei and Lafferty, 2005] and our TWTM, on two corpora, a paper collection in DBLP and a set of movie documents in IMDB, to compare the generalization performance of the three models. In both datasets, we removed the stop words and conducted experiments using 5-fold cross-validation. Figure 2 presents the results on the DBLP dataset. The perplexity of TWTM is larger than or similar to LDA and CTM, when the number of topics $T \leq 100$; when T increases, both CTM and LDA are running into serious over-fitting, while the trend of TWTM keeps going down and the perplexity is significantly lower than those of the baselines. Figure 3 demonstrates the perplexity results on the IMDB data. Clearly, TWTM excels both CTM and LDA significantly and consistently. Figures 2 and 3 show that TWTM works very well in semi-structured document modeling.

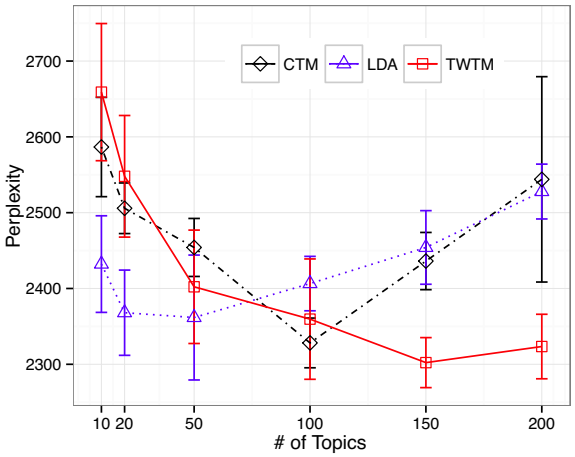

Figure 2: Perplexity results on the DBLP corpus for TWTM, LDA and CTM

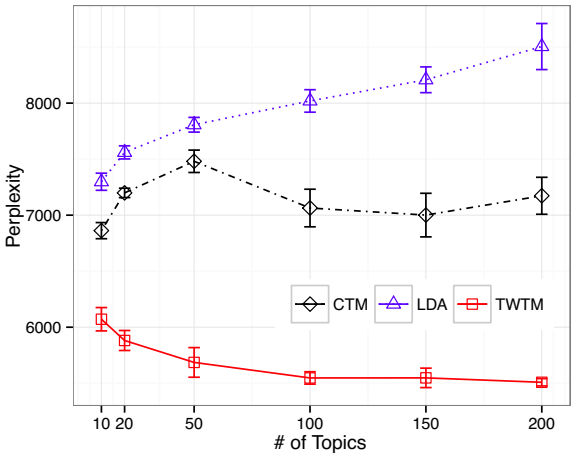

Figure 3: Perplexity results on the IMDB corpus for TWTM, LDA and CTM

4.2 Examples of Tag-Topic Distributions

In the preceding section we demonstrated the performance of TWTM on the learning of tags' distribution by process the New York Times' full text. We treated the 31 days in March 2011 as the tags set of the news corpus, and set the number of topics K = 100. Each topic is illustrated with the top 20 words most likely to be generated conditioned on the topic from the matrix of ψ. We found one topic about the Japan earthquake happened in March 11, 2011. The keyword representation of the topic is shown in Figure 4.

[2]http://www.informatik.uni-trier.de/ ley/db/
[3]http://www.nytimes.com
[4]http://www.imdb.com

japan	earthquake	mercury	messenger	der
van	spacecraft	particles	association	hong
cern	meer	axis	indonesia	plate
weak	orbit	indonesian	halid	opera

Figure 4: An illustration of the topic about the Japan earthquake from the 100-topic solution for the NYTimes collection, including the top 20 words, which TWTM considers are most likely to be generated in the topic.

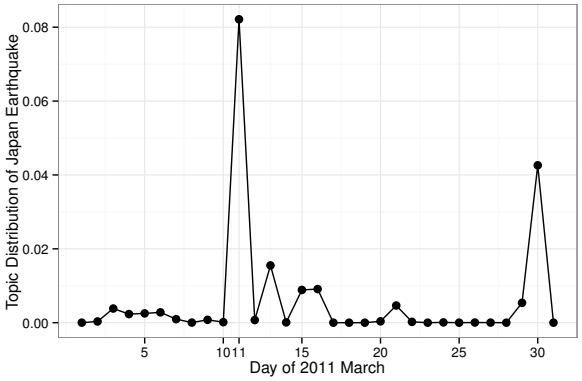

Figure 5: The components of the topic about the Japan earthquake distributed in 31 tags (days of March 2011).

Figure 5 shows that the topic we selected has the different value between the tags, the 31 days. The results shows that at the label of "Mar. 11" the value reaches maximum and there is another extreme value appeared at "Mar. 30". The date of Mar. 11 is when the earthquake happened with the highest value, and at the date of Mar. 30, there are some other news about Japan Nuclear leak , for the example of first detected the radioactive elements plutonium.

The capability of inferring the distributions of tags is obtained directly through the proposed model TWTM. And furthermore, there is no need to distinguish between the different types of labels so as to make use of all the tags observed in the document.

4.3 Classification Performance Comparison

The next experiment is to test the classification performance utilizing feature sets generated by TWTW and other baselines. For the base classifier, we use LIBSVM [Chang and Lin, 2011] with Gaussian kernel and the default parameters. For the purpose of comparison, we trained three SVMs on tf-idf word features, features induced by a 100-topic LDA model, and features generated by a TWTM model with the same number of topics, respectively. Since the TWTM method uses the tag information, we append the tag vector t^d to both the tf-idf and LDA feature sets.

In these experiments, we conducted multi-class classification experiments using the IMDB dataset, which contains 29 genres. We calculated the evaluation metrics recall@1 and recall@3 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 6, where we see that there is significant improvement in classification performance when using LDA and TWTM comparing with tf-idf, and TWTM outperforms LDA in terms of both recall@1 and recall@3.

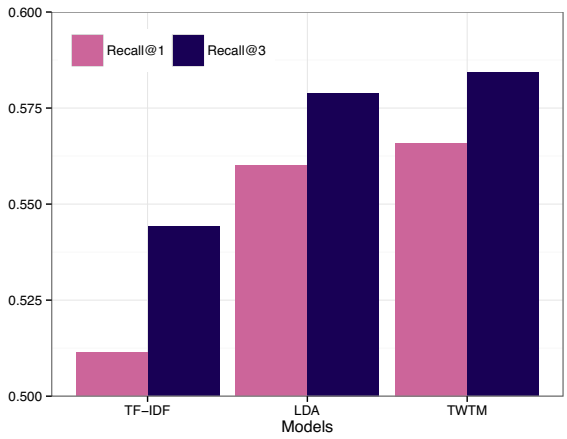

Figure 6: Classification results of TWTM, TF-IDF, and LDA on recall@1 and recall@3 with 5-fold cross-validation.

5 Conclusions

With the tag-weighted topic model proposed in the paper, we provide and analyze a probabilistic approach for mining semi-structured documents. This model provides significantly improved predictive capability in terms of perplexity compared to the other topic models, such as LDA and CTM. Besides, TWTM was shown to be able to obtain the topics distribution of tags in the corpus, which is very useful for text classification, clustering and other data mining applications. Meanwhile, TWTM proposes a novel framework of processing the tagged text with a high extensibility, and uses a novel function of tag-weighted topic assignment of documents. The primary benefit of the tag-weighted topic model is that it allows one to incorporate different types of tags in modeling documents, and provides a general framework for multi-tags modeling at not only the level of tags but also the level of documents. It helps provide different approaches in classification, clustering, recommendation, and so on. In the future, we plan to apply TWTM to different practical areas, especially the problems of the large-scale document modeling.

Acknowledgments

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

References

[Andrieu *et al.*, 2003] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[Asuncion *et al.*, 2009] Arthur U. Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *UAI*, pages 27–34, 2009.

[Attias, 1999] Hagai Attias. A variational baysian framework for graphical models. In *NIPS*, pages 209–215, 1999.

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

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

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

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

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

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

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

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

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

[Deng *et al.*, 2011] 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.

[Griffiths and Steyvers, 2004] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. In *PNAS*, pages 449–455, 2004.

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

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

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

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

[Petterson *et al.*, 2010] 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.

[Ramage *et al.*, 2009] 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.

[Ramage *et al.*, 2011] 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.

[Rosen-Zvi *et al.*, 2004] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494, 2004.

[Sato and Nakagawa, 2012] Issei Sato and Hiroshi Nakagawa. Rethinking collapsed variational bayes inference for lda. In *ICML*, 2012.

[Wei and Croft, 2006] 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.

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