Multi-Label Topic Model Conditioned on Label Embedding

Lin Tang
Key Laboratory of Educational Informatization for Nationalities
Ministry of Education
Yunnan Normal University
Kunming, China
maitanweng2@163.com

Lin Liu
School of Information
Yunnan Normal University
Kunming, China
liulinrachel@163.com

Jianhou Gan*
Key Laboratory of Educational Informatization for Nationalities
Ministry of Education
Yunnan Normal University
Kunming, China
10800662@qq.com

***Abstract*—In most real-world document collections, there are various types of labels that usually carry context information, such as label hierarchies or textual descriptions. Nonetheless, the commonly-used approaches to modeling text corpora ignore this information. Label embedding can reflect more extensive label context information and have a capability of leveraging various sources of information. In this paper, we propose a multi-label topic model conditioned on label embedding, which incorporate label embedding into the generative process of multi-label topic models in text domain, so as to improve documents classification accuracy and topic quality. By introducing Dirichlet-multinomial regression (DMR) framework into a multi-label topic model called Labeled LDA(LLDA), our model apply an exponential priori constructed previously with label embedding on the hyper-parameters of document-label distribution, which reflects the effects of label embedding on label probability distribution. The experimental results demonstrate the potential of our model through an exploration of a standard document dataset.**

Keywords—topic model, multi-label classification, label embedding, hierarchy label

I. INTRODUCTION

With the rapid accumulation of text data in network, more and more studies focus on effective and efficient machine learning techniques to serve the user's access to information and demand. Among these machine-learning techniques, topic model is an important way to discovery complex semantic of documents or other data. As an example, an article can be categorized as either topic 'economy' or topic 'culture', utilizing traditional topic model such as Latent Dirichlet allocation (LDA)[1] for document is able to better reflect true meaning of corpus in the digital humanities, political science, and other related fields [2].

Nevertheless, the latent topic discovered by topic models is mainly used to cluster unlabeled documents. For amount of labeled corpus in microblogs, web forum and so on, it is critical to develop multi-label topic models that facilitate understanding of such labeled data. Therefore, some variations of LDA that focus on text field have been proposed for multi-label classification, which can incorporate label set into topic learning procedure and become a supervised model [3,4]. Among them, LLDA[3] associates each label with a corresponding topic directly, which is one of the easiest and most effective ways to introduce label set into topic model and obtained better results for credit attribution in corpora. Meanwhile, we also found that label sets of text in a network usually carry context information, such as hierarchy relationship information, which can be encoded in label embedding. This data is widespread in question answering platforms, web forums and microblogs. Although existing studies about topic model extending have proposed several hierarchical topic models [5], which only depended on label hierarchical correlation rather than transforming them to label embedding. Label embedding can reflect more extensive label context information and enjoys a built-in ability to leverage alternative sources of information instead of or in addition to attributes, such as class hierarchies or textual descriptions. Therefore, this paper propose a multi-label topic model conditioned on label embedding, which incorporate label embedding into the generative process of multi-label topic models in text domain, so as to improve documents classification accuracy and topic quality.

We first briefly discuss the related work in Section 2. Then, we elaborate on our model in Section 3. The experimental results derived on standard datasets are shown in Section 4. We conclude the article in Section 5.

II. RELATED WORK

In natural language processing, labels embedding for text classification has been studied in several tasks, such as the context of heterogeneous networks [6] and multitask learning [7]. In computer vision, a large number of studies leverage label embedding for image classification [8], multimodal learning [9], and so on. Besides, for the case of zero-shot learning while some classes are unseen, label embedding is able to capture the label correlation to improve the prediction accuracy[10]. In general, Label embedding is able to represent various information of label set and embedded into classifiers. To our knowledge, although label embedding has been studied in several classification model, such as above studies, how to joint multi-label topic modeling and label embedding to make full use of label information for text classification has not been studied previously, representing a contribution of this paper.

Most real-world document collections involve various types of metadata, such as author, source, and date [11]. In our paper, the introduction of label context information into topic model follows the way of document metadata introduction. In recent years, many variants of LDA have been developed to incorporate document and word metadata. At the document level, labels of documents can be considered as the metadata of documents and used to guide topic learning so that more meaningful topics can be discovered. Blei et al. proposed supervised LDA(SLDA) to jointly models words and labels. In reference [12], authors present a topic model called MetaLDA, which is able to leverage either document or word meta-information, or both of them jointly, in the generative process. Especially in reference[13], the DMR model incorporate arbitrary document-level features to inform topic priors. Based on this study, a recent paper also extended DMR by using deep

neural networks to embed metadata into a richer document prior [14].

III. Multi-Label topic model Conditioned on Label Embedding

A. Label embedding and DMR framework

With directed acyclic graph (DAG) label correlation as example, we showed the way of label embedding construction. For a label set $C=\{c_1,c_2,c_3,c_4\}$, there is a DAG correlation between labels, as shown in Fig.1.

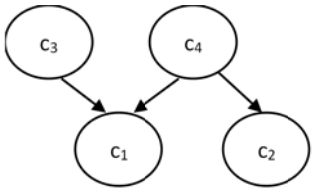

Fig. 1. DAG of label set

We suppose $D=\left[d_{ij}\right]\in\{0,1\}^{4\times4}$ as the label-embedding matrix of above label set. d_{ij} represents the element of D and satisfied: if label c_j is the ancestor of c_i, then $d_{ij}=1$, otherwise 0. So matrix D is shown blow:

$$D=\begin{bmatrix}1&0&1&1\\0&1&1&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

Then, each row of matrix D represents the 'feature' of a label. In other words, D is label embedding. It is important to note that above D is only one way of label embedding, and other label 'feature' can also be constructed as vector for representing label embedding. For these label 'feature', this paper introduces a DMR framework into an LLDA model. In reference [13], DMR framework can incorporate arbitrary document-level features to inform topic priors. Similarly, DMR framework is utilized to model label-level features on prior parameters, which is an upstream topic model. Rather than defining specific random variables in the graphical model for each new label feature, DMR treats the label embedding as features in a log-linear model. The log-linear model parameterizes the Dirichlet prior for the label's word distribution. As DMR framework make no assumptions on model structure, it is flexible to incorporating different types of extra information on prior hyper-parameters.

B. Introducing DMR into label-word parameters

We propose a multi-label topic model conditioned on label embedding (MTM-LE), which extends LLDA model by introducing DMR that learns a transformation of label embedding to form the Dirichlet hyper-parameter.

Suppose there are documents in the training set which compose the document space $\mathbb{D}=\{1,\ldots,D\}$; the global topic space including T topics is represented as $\mathbb{T}=\{1,\ldots,T\}$; the words space is represented as $\mathbb{W}=\{1,\ldots,W\}$; the label space is expressed as $\mathbb{L}=\{1,\ldots,L\}$. According to the definition of LLDA model, there is a one-to-one correspondence between label and topic, then $\mathbb{L}\square\mathbb{T}$, $T=|\mathbb{T}|=L=|\mathbb{L}|$. Besides, the dimension of label embedding space is same as label space $\mathbb{L}=\{1,\ldots,L\}$.

For expressing the observed labels of document, a sparse binary vector $\mathbf{\Lambda}_d=\{\Lambda_{dl}\}_{l=1}^{L}$ is defined as following. In addition to whole label set, $\mathbb{L}_d$ represents the label sub space of document χ_d, then $\mathbb{L}_d\subseteq\mathbb{L}$.

$$\Lambda_{dl}=\begin{cases}1, & l\in\mathbb{L}_d\\0, & l\notin\mathbb{L}_d\end{cases}\qquad(1)$$

In the process of multi-label topic modeling, the whole document set shares label space $\mathbb{L}$, and each label $l\in\mathbb{L}$ in label space corresponds to a multinomial distribution of words space $\mathbb{W}$, whose parameter is $\boldsymbol{\theta}_l=\{\theta_{lw}\}_{w=1}^{W}$, where θ_{lw} is the probability of word w under label l and obey the conjugate prior distribution under hyper-parameter $\boldsymbol{\lambda}_l=\{\lambda_{lw}\}_{w=1}^{W}$. Besides, there is a label features (label embedding) for document χ_d, which can be represented as a feature vector $\mathbf{z}_l=\{z_{li}\}_{i=1}^{L}$. $\boldsymbol{\gamma}_w=\{\gamma_{wi}\}_{i=1}^{L}$ is feature parameters corresponding to label.

According to DMR framework, when label l face a mutually exclusive complete set of options $\mathbb{W}=\{1,\ldots,W\}$, system components of selected utility is $\tilde{\lambda}_{lw}=\exp\left(\mathbf{z}_l\boldsymbol{\gamma}_w^T\right)$, random components is $\zeta_{tw}=\exp\left(\xi_{tw}\right)$.

We suppose ζ_{tw} obey Gamma distribution $\zeta_{lw}\sim\Gamma\left(\delta_l^{-1}\tilde{\lambda}_{lw},\delta_l^{-1}\tilde{\lambda}_{lw}\right)$, where δ_l^{-1} is a positive number related with label l. According to the property of Gamma distribution, we can deduce:

$$\tilde{\lambda}_{lw}\zeta_{lw}=\tilde{\lambda}_{lw}\exp\left(\xi_{lw}\right)\sim\Gamma\left(\delta_l^{-1}\tilde{\lambda}_{lw},\delta_l^{-1}\right)$$

The selected probability under best selected utility is obtained by $\tilde{\lambda}_{tw}\zeta_{tw}$ normalization:

$$\theta_{lw}=\frac{\tilde{\lambda}_{lw}\zeta_{lw}}{\sum_{w=1}^{W}\tilde{\lambda}_{lw}\zeta_{lw}}\sim\Gamma\left(\delta_l^{-1}\tilde{\lambda}_{lw},\delta_l^{-1}\right)$$

Suppose:

$$\lambda_{lw}=\delta_l^{-1}\tilde{\lambda}_{lw}=\delta_l^{-1}\exp\left(\mathbf{z}_l\boldsymbol{\gamma}_w^T\right)$$

We can deduce that $\boldsymbol{\theta}_l=\{\theta_{lw}\}_{w=1}^{W}$ obey Dirichlet distribution of parameter $\boldsymbol{\lambda}_l=\{\lambda_{lw}\}_{w=1}^{W}$:

$$p(\boldsymbol{\theta}_l) = p\left(\{\theta_{lw}\}_{w=1}^{W}\right) = \frac{\Gamma\left(\sum_{w=1}^{W}\delta_l^{-1}\tilde{\lambda}_{lw}\right)}{\prod_{w=1}^{W}\Gamma\left(\delta_l^{-1}\tilde{\lambda}_{lw}\right)}\prod_{w=1}^{W}\theta_{lw}^{\delta_l^{-1}\tilde{\lambda}_{lw}-1} = \frac{\Gamma\left(\sum_{w=1}^{W}\lambda_{lw}\right)}{\prod_{w=1}^{W}\Gamma\left(\lambda_{lw}\right)}\prod_{w=1}^{W}\theta_{lw}^{\lambda_{lw}-1}$$

Therefore, θ_{tw} of DMR framework is the selected probability of individual sample $\mathbf{x}_{dn}$ in label l selecting word option w, and obtain the best selected utility U_{wl}. In LLDA, $\boldsymbol{\theta}_l = \{\theta_{lw}\}_{w=1}^{W}$ is the parameter vector of word multinomial distribution under label l. Then, the hyper-parameter $\boldsymbol{\lambda}_l = \{\lambda_{lw}\}_{w=1}^{W}$ can be represented as:

$$\begin{aligned}\lambda_{lw} &= \delta_l^{-1}\exp\left(\mathbf{z}_l\boldsymbol{\gamma}_w^T\right) = \exp\left(\log\delta_l^{-1}\right)\exp\left(\mathbf{z}_l\boldsymbol{\gamma}_w^T\right)\\ &= \exp\left(\log\delta_l^{-1} + \mathbf{z}_l\boldsymbol{\gamma}_w^T\right) = \exp\left(\log\delta_l^{-1} + \sum_{i=1}^{L} z_{li}\gamma_{iw}\right)\\ &= \exp\left(z_{l,L+1}\gamma_{l,L+1} + \sum_{i=1}^{L} z_{li}\gamma_{iw}\right) \quad = \exp\left(\hat{\mathbf{z}}_l\hat{\boldsymbol{\gamma}}_w^T\right)\end{aligned}$$

Where,

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_w &= \left\{\boldsymbol{\gamma}_w, \gamma_{l,L+1}\right\} = \left\{\gamma_{l1}, \gamma_{l2}, \ldots, \gamma_{lL}, \gamma_{lL+1}\right\}\\ \hat{\mathbf{z}}_l &= \left\{\mathbf{z}_l,\ z_{l,L+1}\right\} = \left\{z_{l1}, z_{l2}, \ldots, z_{lL},\ \ 1\ \right\}\end{aligned}$$

Then, a 'fake observed feature' $z_{l,L+1}=1$ is added to $\mathbf{z}_l = \{z_{li}\}_{i=1}^{L}$. Correspondingly, a feature parameter γ_{lL+1} is also added to $\boldsymbol{\gamma}_w = \{\gamma_{wi}\}_{i=1}^{L}$, the label feature vector $\mathbf{z}_l$ and parameter $\boldsymbol{\gamma}_w$ are expanded to $\hat{\mathbf{z}}_l = \{z_{li}\}_{i=1}^{L+1}$ and $\hat{\boldsymbol{\gamma}}_w = \{\gamma_{wi}\}_{i=1}^{L+1}$ of dimension $L+1$.

C. The generative process

The generative process of our model can be described as follows. The corresponding graphical model is shown in Fig.2.

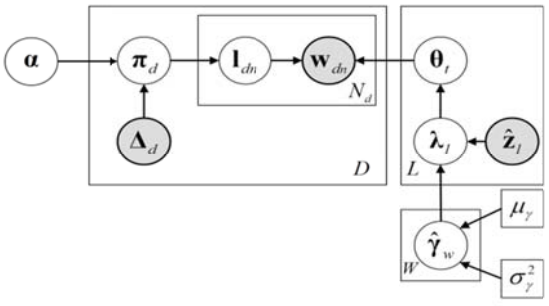

Fig. 2. Graphic model

1) For each global topic $t \in \mathbb{T} = \{1,\ldots,T\}$:

a) Generate label feature weighted parameter vector $\hat{\boldsymbol{\gamma}}_w = \{\gamma_{wi}\}_{i=1}^{L+1}$ of word w from $L+1$ dimensions normal distribution of parameter $\left(\mu_\gamma, \sigma_\gamma^2 I\right)$:

$$\hat{\boldsymbol{\gamma}}_w = \{\gamma_{wi}\}_{i=1}^{L+1} \sim N\left(\mu_\gamma, \sigma_\gamma^2 I\right)$$

b) Compute Dirichlet prior hyper-parameter vector of label-word multinomial distribution parameter $\boldsymbol{\theta}_l$:

$$\boldsymbol{\lambda}_l = \{\lambda_{lw}\}_{w=1}^{W} = \left\{\exp\left(\hat{\mathbf{z}}_l\hat{\boldsymbol{\gamma}}_w^T\right)\right\}_{w=1}^{W}$$

c) Generate multinomial parameter vector $\boldsymbol{\theta}_l = \{\theta_{lw}\}_{w=1}^{W}$ from L dimensions Dirichlet distribution:

$$\boldsymbol{\theta}_l = \{\theta_{lw}\}_{w=1}^{W} \sim \mathrm{Dir}\left(\boldsymbol{\lambda}_l\right)$$

2) For each document $\mathcal{X}_d$, $d \in \mathbb{D} = \{1,\ldots,D\}$:

a) Compute Dirichlet prior hyper-parameter vector of document-label weighted parameter $\boldsymbol{\pi}_d$:

$$\boldsymbol{\alpha}_d = \{\alpha_{dl}\Lambda_{dl}\}_{l=1}^{L}, \quad \Lambda_{dl} = \begin{cases}1, & l \in \mathbb{L}_d\\ 0, & l \notin \mathbb{L}_d\end{cases}$$

b) Generate label weighted parameter vector $\boldsymbol{\pi}_d$ from L dimensions Dirichlet distribution:

$$\boldsymbol{\pi}_d = \{\pi_{dl}\Lambda_{dl}\}_{l=1}^{L} \sim \mathrm{Dir}\left(\boldsymbol{\alpha}_d\right)$$

c) For each word sample $\mathbf{x}_{dn}$:

i. Generate label number $\mathbf{l}_{dn}$ of $\mathbf{x}_{dn}$ from T dimensions multinomial distribution of parameter $\boldsymbol{\pi}_d$:

$$\mathbf{l}_{dn} \sim \boldsymbol{\pi}_d \quad \text{or} \quad \mathcal{L}_d = \{\mathbf{l}_{dn}\}_{n=1}^{N_d} \sim \mathrm{Mul}\left(\boldsymbol{\pi}_d, N_d\right)$$

ii. Generate word number $\mathbf{w}_{dn}$ of $\mathbf{x}_{dn}$ from W dimensions multinomial distribution of parameter $\boldsymbol{\theta}_{\mathbf{l}_{dn}}$:

$$\mathbf{w}_{dn} \sim \boldsymbol{\theta}_{\mathbf{l}_{dn}} \text{ or } \mathcal{W}_d = \{\mathbf{w}_{dn}\}_{n=1}^{N_d} \sim \mathrm{Mul}\left(\boldsymbol{\theta}_{\mathbf{l}_{dn}}, N_{\mathbf{l}_{dn}}\right)$$

As we can see from the generative process in above, $\boldsymbol{\lambda}_l$ is obtained by label embeding vector $\hat{\mathbf{z}}_l$, its weighted parameter is $\hat{\boldsymbol{\gamma}}$. Compared with non-DMR topic model, $\boldsymbol{\lambda}_l$ is a parameter decided by extra information(label embedding), rather than a random variable in traditional multi-label topic model .

In our model, there are four unknown parameters to be estimated: label-word multinomial distribution parameter $\boldsymbol{\theta} = \{\theta_{lw}\}_{l=1,w=1}^{L,\ W}$, the label feature parameters $\hat{\boldsymbol{\gamma}} = \{\gamma_{wi}\}_{w=1,i=1}^{W,\ L+1}$, and label weight $\boldsymbol{\pi}_d = \{\pi_{dl}\Lambda_{dl}\}_{l=1}^{L}$. Besides, $\mathcal{L}_d = \{\mathbf{l}_{dn}\}_{n=1}^{N_d}$ is the hidden variables to be estimated. The observed data are the word samples $\mathcal{W}$ and binary vector $\boldsymbol{\Lambda}$. The joint distribution of $\left(\hat{\boldsymbol{\gamma}}, \boldsymbol{\pi}, \boldsymbol{\theta}, \mathcal{T}, \mathcal{W}\right)$ is shown in below:

$$p\left(\hat{\boldsymbol{\gamma}},\boldsymbol{\pi},\boldsymbol{\theta},\mathcal{L},\mathcal{W}\mid\mu_\gamma,\sigma_\gamma^2,\boldsymbol{\Lambda},\boldsymbol{\alpha},\boldsymbol{\lambda}\right)$$
$$=p\left(\hat{\boldsymbol{\gamma}}\mid\mu_\gamma,\sigma_\gamma^2\right)\cdot\prod_{l=1}^{L}p\left(\boldsymbol{\theta}_l\mid\boldsymbol{\lambda}_l\right)$$
$$\cdot\prod_{d=1}^{D}p\left(\boldsymbol{\pi}_d\mid\boldsymbol{\Lambda}_d,\boldsymbol{\alpha}_d\right)\prod_{n=1}^{N_d}p\left(\mathbf{l}_{dn}\mid\boldsymbol{\pi}_d\right)p\left(\mathbf{w}_{dn}\mid\mathbf{l}_{dn},\boldsymbol{\theta}\right)$$

In general, this kind of method utilize label embedding as prior information of label-word distribution. By gaining more reliable prior distribution for multi-label topic model, so as to reach a more accurate estimation of posterior distributions.

IV. Experiments and Results

A. Dataset and parameter setting

To evaluate and demonstrate the potential of our model, we present a comparative experiment on a standard dataset Reuters. Reuters is a subset of Reuters Corpus Volume I (RCV1) [15], which is a collection of English language stories. Meanwhile, according to subjects of these stories, they are categorized into hierarchical sets. In this dataset, the number of instances is 6000, the number of labels is 77, the number of nodes in the hierarchy is 100, the maximal depth of the hierarchy is 4. Then, the dimension of label embedding is 100.

In our model, there are four different parameters: μ, σ^2, α and λ. α and λ are the parameters of two Dirichlet distribution. We set $\alpha = 50/T$, $\lambda = 200/W$ and $T = L$. μ and σ^2 are the means and variance of normal distribution obeyed by label embedding parameter $\hat{\boldsymbol{\gamma}}$, then we set $\mu = 0$, $\sigma^2 = 1$.

We conduct 10-fold cross validation to measure and compare the performance of our model against LLDA. Three representative multi-label learning evaluation criteria are use in this paper: hamming loss (HL), average precision (AP) and one Error. Besides, we also use three kinds of area under Precision-Recall curve proposed in reference [16], including $\overline{AUPRC}$, $AU(\overline{PRC})$ and $\overline{AUPRC_w}$.

B. Multi-label classification with cross validation

We compare our model with LLDA in Reuters dataset. Table I shows the HL, AP, One Error, $\overline{AUPRC}$, $AU(\overline{PRC})$ and $\overline{AUPRC_w}$ values of these two models in datasets respectively. For AP, $\overline{AUPRC}$, $AU(\overline{PRC})$ and $\overline{AUPRC_w}$, the large the value, the better the performance. Conversely, for HL and One-Error, the smaller the value, the better the performance. It is worth noting that the experimental results of this section are obtained by collapsed Gibbs sampling algorithm.

As shown in Table I, our model achieves the better results in almost all evaluation criteria for Reuters datasets. Concrete analysis is as follows:

TABLE I. The Comparisons of Our Model and LLDA

Method	Multi-label Learning Evaluation Criteria					
	HL	One-Error	AP	$AU(\overline{PRC})$	$\overline{AUPRC}$	$\overline{AUPRC_w}$
LLDA	0.0089	0.02	0.684	0.421	0.1047	0.1175
MTM-LE	0.0074	0.02	0.704	0.472	0.1121	0.1278

As shown in Table I, MTM-LE obtain a better performance in AP, HL, $AU(\overline{PRC})$, $\overline{AUPRC}$ and $\overline{AUPRC_w}$. On HL, MTM-LE achieves 0.0015 reductions over LLDA. On AP, $AU(\overline{PRC})$, $\overline{AUPRC}$ and $\overline{AUPRC_w}$, MTM-LE achieves 0.02, 0.051, 0.0074 and0.0103 improvements over LLDA. Nevertheless, on One-error, MTM-LE gets the same results as LLDA.

Above all, these results indicate that MTM-LE has significant advantages on improving the accuracy of multi-label classification. By introducing DMR framework into LLDA model, the hyper-parameters of label-word weight is optimized from label-embedding, then optimal hyper-parameters rather than fixed setting lead to a better model training results.

V. Conclusions

In this paper, we first investigate label embedding for multi-label classification, and propose the Multi-Label topic model Conditioned on Label Embedding. Based on the label context information, we construct the label embedding as label features. By applying an exponential priori constructed previously with weighted label features on the hyper-parameters of label-word distribution, this model introduce the label embedding into multi-label topic modeling. Finally, MTM-LE is tested on a standard dataset. Compared with LLDA, our model is superior to LLDA. Due to the jointly use of label context information and word information, MTM-LE gets a much more accurate estimation for model posterior distribution.

Acknowledgment

The research is supported by a National Nature Science Fund Project (61562093), Key Project of Applied Basic Research Program of Yunnan Province (2016FA024), Program for innovative research team(in Science and Technology) in University of Yunnan Province, and Starting Foundation for Doctoral Research of Yunnan Normal University (2017ZB013).

References

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[2] Boyd-Graber J, Hu Y, Mimno D. Applications of topic models[J]. Foundations and Trends? in Information Retrieval, 2017, 11(2-3): 143-296.

[3] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL. 2009.

[4] Ramage D, Manning C D, Dumais S T. Partially labeled topic models for interpretable text mining[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011. ACM, 2011.

[5] Liu L, Tang L, He L, et al. An Overview of Hierarchical Topic Modeling[C]// International Conference on Intelligent Human-machine Systems & Cybernetics. IEEE, 2016.

[6] Tang J, Qu M, Mei Q. PTE: Predictive Text Embedding through Large-scale Heterogeneous Text Networks[J]. 2015.

[7] Zhang H, Xiao L, Chen W, et al. Multi-Task Label Embedding for Text Classification[J]. 2017.

[8] Akata Z, Perronnin F, Harchaoui Z, et al. Label-Embedding for Image Classification[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 38(7):1425-1438.

[9] Kiros R , Salakhutdinov R , Zemel R S . Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models[J]. Computer Science, 2014.

[10] Ma Y, Cambria E, Gao S. Label embedding for zero-shot fine-grained named entity typing[C]// Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016: 171-180

[11] Card D, Tan C, Smith N A. Neural Models for Documents with Metadata[J]. 2017.

[12] Zhao H, Du L, Buntine W, et al. Leveraging external information in topic modelling[J]. Knowledge and Information Systems, 2018(3):1-33.

[13] Mimno D. Topic models conditioned on arbitrary features wi th dirichlet-multinomial regression[C]// Proceedings of 24th Conference on Uncertainty in Artificial 1 Intelligence, 2008. 2008.

[14] Benton A, Dredze M. Deep Dirichlet multinomial regression[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 365-374.

[15] Lewis D D, Yang Y, Rose T G, et al. Rcv1: A new benchmark collection for text categorization research[J]. Journal of machine learning research, 2004, 5(Apr): 361-397.

[16] Vens C, Struyf J, Schietgat L, et al. Decision trees for hierarchical multi-label classification[J]. Machine Learning, 2008, 73(2):185-214.