This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

Knowledge Based Topic Model for Multi-modal Social Event Analysis

Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, Changsheng Xu , *Fellow, IEEE*

***Abstract*—With the accumulation of data on the Internet and progress in representation learning techniques, knowledge priors learned from a large-scale knowledge base has been increasingly used in probabilistic topic models. However, it is challenging to learn interpretable topics and a discriminative event representation based on multi-modal information. To address these issues, we propose a knowledge priors- and max-margin-based topic model for multi-modal social event analysis, called the KGE-MMSLDA, in which feature representation and knowledge priors are jointly learned. Our model has three main advantages over current methods: (1) It integrates additional knowledge from external knowledge base into a unified topic model in which the max-margin classifier, and multi-modal information are exploited to increase the number of event descriptions obtained. (2) We mined knowledge priors from over 74,000 web documents. Multi-modal data with these knowledge priors are then incorporated into the topic model to increase the number of coherent topics learned. (3) A large-scale multi-modal dataset (containing 10 events, where each event contained approximately 7,000 Flickr pages) was collected and has been released publicly for event topic mining and classification research. In comparative experiments, the proposed method outperformed state-of-the-art models on topic coherence, and obtained a classification accuracy of 85.1%.**

***Index Terms*—Knowledge Embedding, Multi-Modal, Topic Coherence, Event Classification.**

I. Introduction

With the widespread popularity of social networking sites, social media data are growing rapidly. The development of mobile Internet and digital photography has made it more convenient for people to report events and express their opinions anytime and anywhere, and this has led to an explosion of data on social media sites. When an event occurs (e.g., a football match, an accident on a highway, or an Apple release conference), an increasing amount of multi-media content (e.g., text, image, and video) is uploaded to social media sites by users. Because social event analysis is a type of data-driven process, the huge user-contributed data are extremely helpful and important for obtaining accurate analysis results. However, as users have become the producers and disseminators of data, user-contributed data have become unstructured and noisy, and it is difficult for researchers to use them for social events analysis. Organizing numerous social media data and analyzing social hotspot events automatically are particularly helpful for improving event analysis.

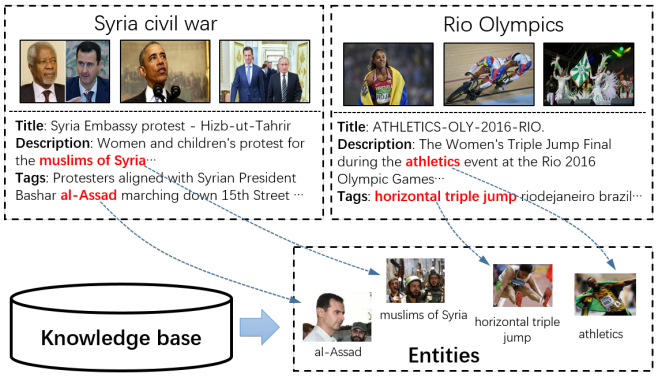


Fig. 1. Demonstration of multi-modal and multi-class properties of social events in Flickr.

Recently, social event analysis has granered the interest of researchers in multi-media analysis, such as for social event detection [1], [2], [3], [4], event tracking [4], and event mining [5]. There are three main challenges in social event analysis.

(1) **A social event is multi-modal.** Traditional social event analysis tasks, such as event tracking and topic mining, involve a single modality, and do not make full use of multi-modal information (e.g., text descriptions and images). On the Internet, social media consists of rich unstructured data with multiple modalities that can complement one another. They help express the complete meaning of social event analysis [6], [7], [8]. For instance, when a new event appears in the web society, different media sites report it from different perspectives. The report text description may be different, but the visual information is often similar.

Considering that past studies have focused on single-modal-based social event analysis [9], [10], [11], it is useful to analyze multi-modal event data in a uniform model.

(2) **A social event involves prior knowledge.** Many entities of social events are freely available in general knowledge bases, and social event data typically contain some important

Feng Xue is with Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education; School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, China.(feng.xue@hfut.edu.cn)

Richang Hong is with Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education; School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, China.(hongrc.hfut@gmail.com). Richang Hong is the corresponding author.

Xiangnan He is with School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei, 230031,China.

Jianwei Wang is with Minglue Technology Group, Room 1002, Block A, Chuangxin Mansion, Haidian District, Beijing, China.

Shengsheng Qian is with National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Road, Beijing, China.

Changsheng Xu is with National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Road, Beijing, China.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

entities, such as people and locations, that may be encoded in these knowledge bases, such as Wikipedia[1] and Freebase[2]. For example, the text in Fig. 1 contains entities (e.g., al-Assad, Muslims of Syria, horizontal triple jump, and athletics) that can be represented as entity pairs in a knowledge base (e.g., a Wikipedia dataset). We consider the must-link relations between these entities in knowledge base as prior knowledge. A knowledge graph is embedded into a low-dimensional continuous vector space in which the inherent structure and certain properties of a large-scale knowledge base are preserved. Specifically, the points indicate entities in vector space, and a single edge for a given pair of entities indicates their relation. Knowledge entity pairs can directly capture the structural information of the original knowledge base, which can be used as a kind of knowledge prior for social event analysis.

(3) **A social event has the multi-class property.** For example, Fig.1 shows two multi-modal social events containing titles, descriptions, tags, and images. Class label information can be used for discriminative feature representation to analyze social events. Therefore, exploiting a model that fuses multi-modal and multi-category information with prior knowledge can help with social event analysis. Many recent studies have explored explored multi-modal information on social media, and proposed methods based on the topic model for social event analysis. For example, Corr-LDA [12] is pioneering research on using relations images of between events and their text descriptions at the topic level. However, these approaches use unsupervised topic models, which cannot use labels to obtain discriminative representations for social event analysis. To address this limitation, some researchers have integrated feature representation learning with discriminative classifiers to obtain a unified classification model [13], [14], [15], [16], [17], [18]. A supervised LDA model was proposed in [18], that uses the continuous response value of linear regression to predict new documents. Wang et al. [19] built a model to represent a discriminative image feature by combining generated topics.

However, the above studies use only softmax regression to relate supervised information with data representations, which yields suboptimal classification results. Moreover, some traditional methods without prior knowledge often generate topics that are difficult to interpret. Topic model-based knowledge is a new research area that has potential for use in event mining. The relevant methods exploit the low-dimensional continuous vector of entities in the knowledge graph, which helps retain the internal structure and certain properties of a large-scale knowledge base. Andrzejewski et al. [20] proposed topic-in-set knowledge to add partial supervision to latent Dirichlet allocation(LDA). The authors of Ref. [21] proposed a fold-all framework to extend topic-in-set knowledge in which first-order logic is used to specify general knowledge. In [22], human knowledge and the topic model were integrated, and a Probase-LDA on probabilistic knowledge was designed to boost topic coherence performance. Although current models combine prior knowledge and the topic model using different approaches, they do not consider large-scale triple-oriented knowledge graphs.

We propose a knowledge and max-margin based topic model (KGE-MMSLDA) for multi-modal event classification and analysis, where the topic model and knowledge entity priors are combined. Our proposed KGE-MMSLDA explicitly models multimedia documents with knowledge that is automatically learned from a pre-existing knowledge graph. In Section IV, experimental results show that KGE-MMSLDA extracts more coherent topics than other compared methods. Moreover, we introduce the max-margin classifier as a regularization term to the topic model. Specifically, the max-margin classifier and KGE-MMSLDA model are integrated into a united generative model that is more conducive to the optimizing model of event analysis. For feature representation, we design multi-modal feature representation where prior knowledge is integrated into the training process of the latent topic relevance. In the classifier design, the max-margin classifier is used for social event classification because it is a powerful discriminant classifier. Specifically, Gibbs sampling is used in the training of the KGE-MMSLDA. We empirically evaluated our proposed model through topic coherence and social event classification on an empirical dataset. Both qualitative and quantitative results show its effectiveness. Our study makes the following contributions:

- A unified framework(KGE-MMSLDA) is proposed where knowledge priors and a max-margin topic model are integrated for multi-modal social event analysis.
- The KGE-MMSLDA model uses prior knowledge mined from external knowledge base and supervised information as part of the regularization to output a reasonable representation and coherently mined topics for social event classification.
- No public dataset is available for multi-media analysis research. We collected a large-scale multi-media dataset called HFUT-mmdata[3] for public social event analysis research. We conducted experiments on it to verify that our KGE-MMSLDA outperforms current topic models.

The remainder of this paper is organized as follows: We first discuss related work in Section II and present the details of KGE-MMSLDA and its optimization process in Section III. We report the experimental results in Section IV. In Section V, we offer the conclusions of this study and highlight avenues for future work in the area.

II. Related Work

Social Event Analysis: Social event analysis has attracted considerable attention from researchers in multimedia analysis. A well-designed event classification model can help boost the performance of event classification tasks. In general, a social event analysis algorithm consists of two processes: feature representation and classifier training. Early event classification methods considered these processes as separate [23], [24]. Blei et al. [23] used the LDA to learn the text representation

[1]https://dumps.wikimedia.org/

[2]https://en.wikipedia.org/wiki/Freebase

[3]http://scholarhub.cn/ScholarHubProject/MMTM/HFUT-mmdata.zip, HFUT stands for HeFei University of Technology

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

for text classification. Qian et al. proposed a multi-modal event topic model, that models social media documents and learns text image correlations to separate topics in text representation from those not in it [25]. Gao et al. [26] proposed a deep learning-based approach to perform for event classification using microblogs. The above methods mainly focused on event feature design, that is used to train the classifier separately. However, they do not integrate event feature representation and classifier training into a unified model. Many supervised models have been proposed to improve social event classification by introducing a classifier to the topic model [15], [16], [18], [27].

Simon et al. [16] proposed a model called the DiscLDA that applies a class-related linear transformation to mixed topic proportions. Blei et al. [18] introduced a supervised LDA (sLDA) to associate categorical information with each document that jointly models documents and supervised information. In [28], two supervised topic models were proposed to solve event classification and regression problems. They account for heterogeneity and biases among different annotators that are encountered in practice when learning from crowds.

The above supervised methods calculate the probability that a document belongs to each event class using maximum likelihood estimation, and the classification is regarded as a process of maximum voting that ignores the powerful discriminative classification algorithm. Recently, researchers have noticed this problem, and some powerful discriminative classifiers have been exploited for social event classification [19], [29], [30]. Wang et al. [19] improved the LDA model using a max-margin classifier for the image classification task. Zhu et al. [29] proposed an alternative approach by considering a new max-margin loss to jointly model the max-margin classifier and feature representation learning for multi-class and multi-label text classification. However,the max-margin methods in [19], [29] focus on text or images separately, where multi-modal information in social media data is often ignored. Unlike these models, in this study, supervised information and multi-modal information of social events is unified in a framework in which multi-modal event representation is learned using a discriminative max-margin classifier.

Topic Model: Topic models have been researched for many years, and are proved to be effective for modeling social events.

By learning document representation, the traditional LDA [23] solves the problem of topic mining and text categorization. Many improvements on the proposal in [23] have been made, such as [13], [17], [18], [31] and [32].Bao et al. [17] proposed a partial sLDA model for cross-domain learning, and in [33], a hierarchical sLDA was presented for a hierarchical multi-media data structure.

The above methods are useful for mining document information for social events. However, only text information is considered in these methods, and other modal information has been ignored. To address the above issues, researchers have proposed methods that make full use of the multi-modal properties of documents[34], [35]. Prabhudesai et al. proposed an extended LDA topic model where the topics are represented using the Gaussian mixture models [36]. In [34], an improved multi-modal LDA model using a topic regression technique was designed to capture correlations between visual features and annotation text for image and video annotation tasks. In [37], Qian et al. proposed a multi-modal sLDA (MMsLDA) for social event classification that can jointly learn textual and visual topics across multi-modal social media data.

Recently, researchers have attended to topic models with domain knowledge to clarify the interpretation of the generated topics. The work in Ref. [38] develops a knowledge-based topic model by incorporating knowledge graph embedding into it for topic mining and document classification. In a similar spirit, our KGE-MMSLDA model combines knowledge entity priors and the multi-modal topic model to jointly learn event representation and the classifier in a unified framework. The work in Ref. [20] first combines domain knowledge with the LDA model and then uses the Dirichlet forest prior to encoding the two basic types of domain knowledge (must-links and cannot-links). Knowledge graph embedding is a new research area in which the entities and relationships of a knowledge graph are embedded into continuous vectors [39], [40], [41], [42], [43]. To automatically learn word correlation knowledge topic modeling with automatically generated must-links and cannot-links (AMC) was proposed in [44] to improve topic modeling in each domain. To achieve the goal of interpretation, [45] proposed a hierarchical topic model, called the graph-sparse LDA model using "controlled structured vocabularies." The work in [22] designed an integrated model, called Probase-LDA, where probabilistic knowledge and the topic model are combined to mine useful topics from multi-media documents. In [46], an extended multi-nomial model was proposed to improve performance, where latent embedding vectors were applied to a large corpus. By replacing the traditional multi-nomial distribution of the LDA model with the Gaussian distribution, [47] used Euclidean distance to calculate the similarity among word vectors to capture semantic regularities in language.

Visual Feature Extraction: The development of deep neural networks (DNNs) has made them prominent in computer vision. Krizhevsky et al. [48] proposed a large and deep convolutional neural network to classify images in the ImageNet dataset and won first place in the ImageNet LSVRC-2012 contest. In [49], an architecture with very small convolution filters and very deep networks was designed, where this demonstrated the importance of depth in visual representations. Recently, most researchers have focused on ways to increase the depth of convolution networks. However, the increasing depth of networks leads to gradients vanishing or exploding, which makes the networks difficult to train. In [50], He et al. proposed a deep residual learning framework with a depth of up to 152 layers that was easy to optimize and achieved good performance. Because of the efficiency of the residual connection, Gao et al. [51] introduced direct connections to every layer in the network and proposed an architecture called DenseNet that can achieve state-of-the-art performance with few parameters and little computation. The work in [52] focused on the relations between feature maps and proposed the "squeeze-and-excitation" block to capture the interdependencies between channels that can be simply

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

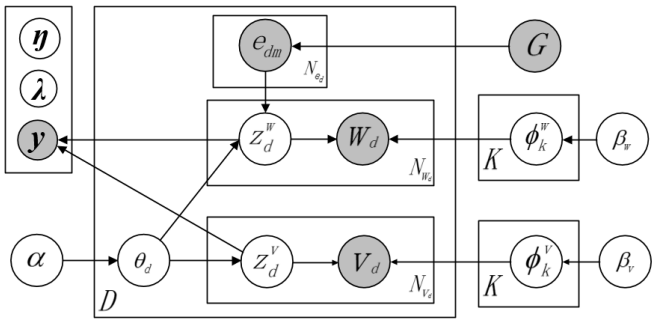


Fig. 2. Framework of the KGE-MMSLDA model for social event classification.

integrated into CNN architectures to yield better performance.

Our KGE-MMSLDA model is a generic framework that incorporates the aforementioned visual feature extraction techniques. In our implementation, for computational efficiency, AlexNet in [48] is used to generate visual feature vectors.

III. Proposed Model

In this section, we define the problem to tackle here and describe the training of our proposed model.

A. Problem Definition

It is important to be able to automatically analyze a social event because multi-media data on the Internet are massive in size and confusing in their variety. During the collection of our HFUT-mmdata dataset, a number of social events are first defined, and images and text belonging to each event were downloaded from well-known websites, such as Flickr. We consider the image and its text description to be a document. In this study, two modalities, image and text, were used for learning.

To depict multi-media information, we used a traditional bag-of-words model to represent the text [23], and a similar data structure was used for images.

A document presentation was converted into a word count vector, where the word order and its context cues were ignored. Let $D = \{(d_1, y_1), (d_2, y_2), \ldots, (d_M, y_M)\}$ be a set of multi-media documents, where d denotes each document that consists of two modalities $d = \{w, v\}$, where w and v, represent the text and image inputs, respectively and y_i denotes the event class number of a given document. Its value is in the range $[1, L]$, where L denotes the number of event classes. The proposed KGE-MMSLDA is a general framework that can be expanded conveniently to exploit data from more modalities. Moreover, the knowledge base G is introduced to the KGE-MMSLDA a priori to improve performance.

We detail the framework of the KGE-MMSLDA model in the following subsection:

B. Unified Framework of KGE-MMSLDA

The framework of the proposed KGE-MMSLDA model has two aspects:(1) a knowledge-based topic model that learns a probabilistic distribution over latent topics for the given documents with prior knowledge, and (2) a delicate classifier in which max-margin theory is used and supervised information is considered during the training phase.

1) KGE-MMSLDA Model: We propose a knowledge-based topic model that is designed to mine latent topics from multi-modal data and learn the representation of a multi-media event using probabilistic distributions over the found topics. In the KGE-MMSLDA, knowledge entity pairs and multiple modal information are incorporated into a unified model. All text words, image words, and entities are sampled in the same topic space. Specifically, documents that consist of multi-modal data and knowledge entity priors are mined from a large-scale knowledge base.

Typically, most documents on the Internet consist of both visual content and corresponding text. To better represent documents and learn text image correlations, we propose a multi-modal topic model that mines correlations between visual and textual contents. Moreover, the proposed model incorporates knowledge priors into multi-modal topic modeling by taking knowledge entities as another input to the model to guide the traditional topic model to discover more coherent topics.

Knowledge entity embedding is a useful method to embed entities and relations inside the knowledge base into continuous vectors where original knowledge is preserved. Each entity is considered as a high-dimensional point in the vector space and each relation as an operation over the entities. When sampling an entity, the model automatically searches for the entity from the knowledge base. For example, when the word "Trump" is sampled, the model obtains the corresponding entity embedding from the knowledge base in which this entity may be spatially similar to the topic "politics." Recent advances in modeling continuous entity embeddings [38] have shown that entity embedding lies on a unit sphere and vMF distribution to model it instead of a multivariate Gaussian distribution [53] makes the inference much more efficient. The probability density function of the vMF distribution is

$$f(x|\mu, \kappa) = C_l(\kappa) \exp\left(\kappa \mu^T x\right) = \frac{\kappa^{0.5l-1}}{(2\pi)^{0.5l} I_{0.5l-1}(\kappa)} \exp\left(\kappa \mu^T x\right) \tag{1}$$

where $x \in \mathbb{R}^l$ lies on an $l-1$ dimensional sphere, u is the average vector and κ denotes the inverse of the variance of the training set. $I_a(b)$ denotes the modified Bessel function of the first type for order a and argument b.

In this article, the result of our knowledge mining is a set of word pairs $< w1, w2 >$, which are frequently co-occurred. During the sampling phase of textual words, we simply set the topic assignment of $w1$ to $w2$ since the two words often appear together.

Fig. 2 shows the graphical representation of the KGE-MMSLDA, where the shaded nodes are observed variables and unshaded nodes are latent variables.

Let $E = \{d_1, d_2, \ldots, d_M, G\}$ be a collection of event documents, where M is the number of input documents. For each multi-media document $d = \{w_d, v_d\}$, w_d, and v_d denote textual and visual words, respectively. $G = \{e_1, e_2, ..., e_{N^G}\}$ denotes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

the set of knowledge entities, e_{dm} denotes the embedding of the m-th entity in document d, and N^G denotes the number of entities. The KGE-MMSLDA can be considered a two-layer model with a topic layer and a word layer. In the topic layer, topic k is considered a K-dimensional multi-nomial distribution of document d. In the word-layer, textual word w and visual word v are considered D_w-dimensional and D_v-dimensional multi-nomial distributions, respectively, of topic k, where D_w and D_v are the respective sizes of the textual and visual vocabulary.

Accordingly, in the KGE-MMSLDA, we generate document d using the following steps:

1. For document d, draw $\theta_d|\alpha \sim Dir\left(\alpha_{\theta_d}\right)$.
2. For visual topic $k \in \{1,2,...,K\}$, draw $\phi_k^v|\alpha_{\phi_k^v} \sim Dir\left(\alpha_{\phi_k^v}\right)$.
3. For each textual topic $k \in \{1,2,...,K\}$,
   (1) draw $\phi_k^w|\alpha_{\phi_k^w} \sim Dir\left(\alpha_{\phi_k^w}\right)$;
   (2) draw $\mu_k \sim vMF\left(\mu_0, C_0\right)$; and
   (3) draw $\kappa_k \sim \log Normal\left(m, \sigma^2\right)$.
4. For visual word v_d in document d,
   (1) draw a topic $z_d^v|\theta_d \sim Mult\left(\theta_d\right)$; and
   (2) draw a word $v_d|z_d^v, \phi^v \sim Mult\left(\phi_{z_n}^v\right)$.
5. For textual word w_d in document d,
   (1) draw a topic $z_d^w|\theta_d \sim Mult\left(\theta_d\right)$; and
   (2) draw a word $w_d|z_d^w, \phi_w \sim Mult\left(\phi_{z_n}^w\right)$.
6. Draw class label $y_d|z_d \sim \max -m\arg in\left(\bar{z}_d, \eta\right)$

In the above, θ_d denotes the distribution of topics in the documents, ϕ_k^v and ϕ_k^w denote the multi-nomial distributions of topic-specific k over visual word v and textual word w, η denotes class coefficients learned by the max-margin classifier, $\bar{z}_d$ denotes the empirical ratio of topics that appear in the textual or visual modality to the total number of topics in the given document d.

As shown in Fig. 2, textual topic z_d^w and visual topic z_d^v are sampled from the same topic distribution $\theta(d)$. In the generative process of a document, text and image information are processed independently, and the topic distribution(θ) and word distribution(ϕ) are updated consequently. For the event class label, we use the max-margin classifier in the KGE-MMSLDA to jointly learn feature representation.

For textual input, our KGE-MMSLDA model considers the learning phase of the parameters in event class c as a single task; thus, there are L tasks, where L is the number of event types. Similarly, there are L tasks for the visual input. Consequently, there are $I = 2 \cdot L$ tasks in the KGE-MMSLDA, and the linear discriminative function of each task $i, i \in I$ is defined as:

$$F_i(\eta_i, \mathbf{z}_i; \mathbf{w}_i, \mathbf{v}_i) = \eta_i^T \bar{\mathbf{z}}_i \tag{2}$$

where $\mathbf{z}$ represents feature vectors of the topic assignment used to couple the topic model and max-margin classifier.

The probability of the max-margin classifier in the topic model is similar to that in [29]. As in that method, we use η as random variables to perform Bayesian estimation by transforming the classifier into a probabilistic distribution, which is formulated as:

$$\begin{aligned}\varphi_i\left(y_d^i|\mathbf{z}_d, \eta\right) &= \exp\left(-2c\max\left(0, T - y_d^i \eta_i^T \bar{z}_d^i\right)\right) \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_d^i}} \exp\left(-\frac{\left(\lambda_d^i + c\zeta_d^i\right)^2}{2\lambda_d^i}\right) d\lambda_d^i\end{aligned} \tag{3}$$

where λ denotes the augmented variables and i is the index of the learning task. According to the above definition, the normalized posterior distribution of the KGE-MMSLDA is formulated as

$$\begin{aligned}&q\left(\eta, \lambda, \mathbf{z}, \theta, \phi_w, \phi_v\right) \\ &= \frac{p_0\left(\eta, \mathbf{z}, \theta, \phi_\mathbf{w}\right) p\left(\mathbf{w}, \mathbf{v}, \mathbf{e}|\mathbf{z}, \phi_\mathbf{w}, \phi_\mathbf{v}\right) \varphi\left(\mathbf{y}, \lambda|\mathbf{z}, \eta\right)}{\psi\left(\mathbf{y}, \mathbf{w}, \mathbf{v}, \mathbf{e}\right)}\end{aligned} \tag{4}$$

where $\psi\left(\mathbf{y}, \mathbf{w}, \mathbf{v}, \mathbf{e}\right)$ is a normalized constant. $p_0\left(\eta, \mathbf{z}, \theta, \phi_\mathbf{w}\right)$ is the prior distribution, $p\left(\mathbf{w}, \mathbf{v}, \mathbf{e}|\mathbf{z}, \phi_\mathbf{w}, \phi_\mathbf{v}\right)$ is the conditional probability of the generation process, and $\varphi\left(\mathbf{y}, \lambda|\mathbf{z}, \eta\right)$ is the posterior distribution that indicates the category information.

2) Model inference: In this part, collapsed Gibbs sampling is used to infer the joint posterior probability to sample the latent topic assignments z_d^w, z_d^v in the KGE-MMSLDA. The basic idea of collapsed Gibbs sampling is to integrate out the model parameters as a joint probability distribution. During sampling process, the new values of the latent variables are iteratively updated according to the previous states of the conditional distributions.

The conditional probabilities of latent variables $\mathbf{z_d^w}$, $\mathbf{z_d^v}$, η, and λ are formulated as below and the derivations of $\mathbf{z_d^w}$ and $\mathbf{z_d^v}$ are detailed in the Appendix.

z_d^w **sampling process**:

$$\begin{aligned}&p\left(z_d^w = k|z_{\neg(d)}^w, \mathbf{w}, \mathbf{v}, \mathbf{e}, \eta, \lambda\right) \\ &\propto \frac{\left(n_{d,k}^{\neg(w)} + \alpha\right)}{\sum_{k=1}^K \left(n_{d,k}^w + \alpha\right)} \frac{\left(n_{k,t}^{\neg(w)} + \beta_w\right)}{\sum_{t=1}^{D_w}\left(n_{k,t} + \beta_w\right)} \cdot \\ &\prod_{i=1}^L \exp\left(\frac{y_d^i c(cT + \lambda_d^i)\eta_{i,k} - \frac{2N_d^w - 2}{N_d^w}\Lambda_{d,n}^i \eta_{i,k} - \frac{c^2}{2N_d^w}\eta_{i,k}^2}{N_d^w \lambda_d^i}\right)\end{aligned} \tag{5}$$

where $n_{d,k}^{\neg(w)}$ is the number of textual words assigned to the latent topic k of document d, except the current textual word w, $n_{d,k}^w$ is the number of textual words assigned to the latent topic k of document d, $n_{k,t}^{\neg(w)}$ is the number of textual words t of topic k appearing in all documents excluding the current word w. $n_{k,t}$ is the number of textual words t of topic k that appear in all documents.α and β_w are priors of the Dirichlet distribution and

$\Lambda_{d,n}^i = \frac{1}{N_d - 1} \sum_{k=1}^K \eta_{i,k} n_{d,\neg(n)}$ denotes the discriminant function value of the words excluding the current word n.

As shown in (5), each task i influences each latent topic z_d^w, which means that all the feature modalities are influenced by one another.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

z_d^v **sampling process**:

$$p\left(z_d^v = k | z_{\neg(d)}^w, \mathbf{w}, \mathbf{v}, \mathbf{e}, \eta, \lambda\right) \propto \frac{\left(n_{d,k}^{\neg(v)} + \alpha\right)}{\sum_{k=1}^{K}\left(n_{d,k}^{v} + \alpha\right)} \frac{\left(n_{k,t}^{\neg(v)} + \beta_v\right)}{\sum_{t=1}^{D_v}\left(n_{k,t} + \beta_v\right)} \cdot \prod_{i=1}^{L} \exp\left(\frac{y_d^i c(cT + \lambda_d^i)\eta_{i,k} - \frac{2N_d^v - 2}{N_d^v}\Lambda_{d,n}^i \eta_{i,k} - \frac{c^2}{2N_d^v}\eta_{i,k}^2}{N_d^v \lambda_d^i}\right) \quad (6)$$

where $n_{d,k}^{\neg(v)}$ is the number of visual words assigned to the latent topic k of document d excluding current visual word v, $n_{d,k}^{v}$ is the number of visual words of topic k of document d; $n_{k,t}^{\neg(v)}$ is the number of visual words t of topic k excluding the current visual word v that appear in all documents, $n_{k,t}$ is the number of visual words t of topic k that appear in all documents and α and β_v are priors of the Dirichlet distribution.

z_d^e **sampling process**:

$$p\left(z_{dn}^e = k | z_{\neg(dn)}^e, w, v, e_{\neg(dn)}, \alpha, \mu_0, C_0, m, \sigma\right) \propto \frac{\left(n_{d,k}^{\neg(e)}\right)}{\sum_{k=1}^{K}\left(n_{d,k}^{e}\right)} \cdot C_L(\kappa_k) \frac{C_L\left(\left\|\kappa_k \sum_{i \neq dn, z_i^e = k} e_i + C_0\mu_0\right\|\right)}{C_L\left(\left\|\kappa_k \sum_{i: z_i^e = k} e_i + C_0\mu_0\right\|\right)} \quad (7)$$

$$p\left(\kappa_k | \kappa_{\neg k}, \ldots\right) \propto \frac{C_L(C_0) C_L(\kappa_k)^{n_k^e}}{C_L\left(\left\|\kappa_k \sum_{i: z_i^e = k} e_i + C_0\mu_0\right\|\right)} \cdot \log Normal\left(\kappa_k | m, \sigma^2\right) \quad (8)$$

where $z_{\neg(dn)}^e$ denotes the topic assignments for all entities except e_{dn}; $e_{\neg(dn)}$ denotes the embedding of all entities, except e_{dn}; $n_{d,k}^{\neg(e)}$ denotes the number of entities assigned to latent topic k except the current entity e, and κ_k is drawn from the log-normal distribution. We first sample κ_k samples from $\log Normal\left(\kappa_k | m, \sigma^2\right)$, and then sample the final κ_k from them. μ_0, C_0, m, and σ denote the hyper-parameters that control the corresponding vMF parameters μ_k and κ_k.

η_i **sampling process**:

$$p(\eta | \mathbf{z}, \lambda) = \prod_{i=1}^{I} p(\eta_i | \mathbf{z}_i, \lambda_i) = \prod_{i=1}^{I} N\left(\eta_i; \mu_i^\eta, \Sigma_i^\eta\right) \quad (9)$$

where the mean matrix and covariance matrix are written as follows:

$$\mu_i^\eta = \Sigma_i^\eta \left(c \sum_{d=1}^{D} y_d^i \frac{\lambda_d^i + cT}{\lambda_d^i} \bar{\mathbf{z}}_d^i\right) \quad (10)$$

$$\Sigma_i^\eta = \left(\frac{1}{\sigma^2} I + c^2 \sum_d \frac{\bar{\mathbf{z}}_d^i \bar{\mathbf{z}}_d^{i,(\mathrm{T})}}{\lambda_d^i}\right)^{-1} \quad (11)$$

λ_d^i **sampling process**:

$$p\left(\lambda_d^i | \mathbf{z}_d^i, \eta\right) \propto \frac{1}{\sqrt{2\pi\lambda_d^i}} \exp\left(-\frac{\left(\lambda_d^i + c\varsigma_d^i\right)^2}{2\lambda_d^i}\right) \quad (12)$$

$$= GIG\left(\lambda_d^i; \frac{1}{2}, 1, c^2(\varsigma_d^i)^2\right) \quad (13)$$

where $GIG(x; p, a, b)$ is the generalized inverse Gaussian distribution, and $GIG(x; p, a, b) = C(p, a, b) x^{p-1} \exp\left(-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right)$.

C. Classification of a Multi-media Social Event

Following Gibbs sampling, we can estimate the parameters ϕ^w,ϕ^v and θ_d of the textual and visual modalities as in [54].

$$\phi_{k,t}^w = \frac{n_{k,t}^w + \beta_w}{\sum_{p=1}^{D_w}\left(n_{p,k}^w + \beta_w\right)} \quad (14)$$

$$\phi_{k,t}^v = \frac{n_{k,t}^w + \beta_v}{\sum_{p=1}^{D_v}\left(n_{p,k}^v + \beta_v\right)} \quad (15)$$

$$\theta_{d,k}^w = \frac{\left(n_{d,k}^w + \alpha\right)}{\sum_{k=1}^{K}\left(n_{d,k}^w + \alpha\right)} \quad (16)$$

$$\theta_{d,k}^v = \frac{\left(n_{d,k}^v + \alpha\right)}{\sum_{k=1}^{K}\left(n_{d,k}^v + \alpha\right)} \quad (17)$$

As formulated above, we can obtain updated parameters ϕ^w, ϕ^v, $\theta_{d,k}^w$, $\theta_{d,k}^v$, and η. Given a new event document d_{new} composed of textual words w_{new} and visual words v_{new}, we first sample topic distributions $\theta_{d,k}^w$ and $\theta_{d,k}^v$ of the document. We then obtain the average topic assignment vector of the new document $\bar{\mathbf{z}}_{new}$ and parameters η after the Gibbs sampling phase. Ultimately, the category of the new document is predicted based on (2). Based on these predictions, we use a maximum majority voting to obtain the final category of the new document.

IV. Experimental Results

We performed experiments on a publicly accessible dataset to verify the effectiveness of our proposed method. We first describe the dataset structure, detail the feature extraction process, and finally present the results and analysis.

A. Experimental Settings

Considering that the social analysis community does not have a mature multi-modal public dataset for topic mining and event classification, We created a dataset called HFUT-mmdata for social event analysis research using content from Flickr. We exploited the official API provided by Flickr to crawl images and the corresponding textual information, such as titles, descriptions and tags, for predefined events. The dataset contained 74,364 documents belonging to 10 types of events with approximately 7,000 to 9,000 documents inside each event. Each event, contained two types of modal information: text and images. The dataset was thus suitable for multi-modal social event analysis. It contained unrelated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

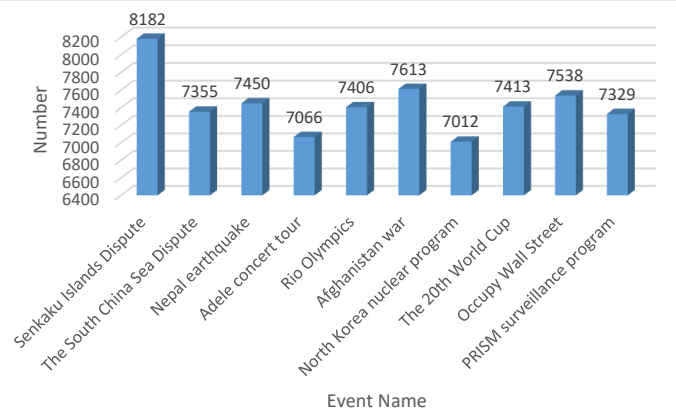


Fig. 3. The statistics of our social event dataset(HFUT-mmdata).

events (e.g., "Adele concert tour," "Nepal earthquake," and "PRISM surveillance program") as well as similar events (e.g., "Senkaku Islands dispute" and "South China Sea dispute").

The details of the 10 social events are listed in Fig. 3.

B. Implementation Details

Feature Extraction: For text data, we first performed pre-processing (removed stop words and stemming) and then simply extracted the bag-of-words features of the input text. We extracted textual feature vectors by counting the word frequency; that is, each entry in the textual feature vector denoted the number of occurrences of a word. For image data, we first divided each image into N patches (e.g., 25 patches) on average resized them to a fixed size, and used AlexNet to extract a 4,096-dimensional feature vector from each patch. We used PCA to reduce them to a low-dimensional space. Finally, the vectors of all image patches were clustered into 5,000 clustering centers by using k-means classification. Thus, each image was converted into N words (a patch corresponded to a visual word) of the 5,000 visual center words. Ultimately, these images were processed using the topic model, in the same manner as textual information.

External Knowledge: WordNet[4] (or its subset, WN18[5], introduced in [40]), is a well-known large-scale lexical database in which words are categorized into sets of cognitive synsets, and is often used by researcher for external knowledge. In our implementation, we used the documents in HFUT-mmdata as corpus to mine knowledge priors. To verify the superior performance of the KGE-MMSLDA, we compared it with following baselines most related to it:

- LDA [23]: This traditional LDA model mines valuable topics from documents. We implemented the LDA algorithm on textual information in our HFUT-mmlda dataset.
- MMLDA[55]: This method models multi-modal data based on the traditional LDA to obtains more meaningful topics .Textual and visual words are counted in this model.
- Link-KGE-LDA[38]: This is a knowledge-based topic model, which integrates knowledge graph embedding with the traditional topic model. This model introduces additional knowledge into topic models using an entity vector. In addition to sampling text words, this model also samples entity words in the training process.
- Corr-KGE-LDA[38]: Similar to Link-KGE-LDA, this model is a knowledge-based model, which integrates knowledge graph embedding into a traditional topic model in the same manner. The difference is that entity embedding is constructed from the topics of the words that appear in the same document during the sampling process.

[4]https://wordnet.princeton.edu/

[5]https://everest.hds.utc.fr/doku.php?id=en:transe

C. Results and Analysis

In this section, we first present the evaluation of our method and the baselines on topic mining in terms of the topic coherence index and then assess their performance in terms of event classification. Finally, we present a qualitative evaluation of the multi-modal topics mined by our model.

1) Coherence of Topic Mining: Topic coherence has been regarded as an efficient measurement of the human ability to interpret the mined topics [56], [57]. To evaluate the quality of the topics mined by each model, we used pointwise mutual information (PMI) [58] to measure topic coherence.

We used the PMI [58] to assess the topic coherence of each model:

$$PMI(k) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{p(w_i, w_j)}{p(w_i)\, p(w_j)} \tag{18}$$

where N denotes the top word number of topic k and $p(w_i)$ denotes the probability that word w_i occurs in a document. $p(w_i, w_j)$ denotes the probability that w_i and w_j occur in the same document. Similar to most methods [38], a higher PMI score in ours signified a more coherent topic. Therefore, the top 10 words of each topic were used to calculate the PMI score. Moreover, 4,776,093 English Wikipedia articles [6] were used to compute the PMI score.

The number of output topics K was a hyperparameter. To choose the best value of K in the experiment, we calculated the PMI scores of the KGE-MMSLDA from $K = 10$ to $K = 100$. From Fig. 4, we see that the PMI scores increased rapidly from $K = 10$ to $K = 40$, and the growth rate slowed and eventually stabilized from $K = 50$ to $K = 100$. In general, a larger value of K leads to longer training, whereas a smaller value of K worsens performance. Therefore, we fixed $K = 40$ to compare the PMI scores of the baselines. For other datasets, the balance point of K should be obtained experimentally.

Table I shows the PMI scores of all models run on our HFUT-mmdata dataset. Based on these results, we can draw the following conclusions:

(1) The MMLDA outperformed the traditional LDA, which shows that multi-modal information can improve the interpretation of the topics.

(2) Link-KGE-LDA and Corr-KGE-LDA were better than traditional LDA and MMLDA, which means that the topic model with knowledge priors improved topic interpretability significantly.

[6]http://deepdive.stanford.edu/opendata/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

TABLE I
THE PMI SCORES COMPARED WITH OTHER EXISTING METHODS.

Methods	PMI scores
LDA	78.1
MMLDA	79.8
Corr_KGE_LDA	79.7
Link_KGE_LDA	82.5
KGE-MMSLDA	**89.7**

(3) Our KGE-MMSLDA outperformed traditional LDA and MMLDA model by 9-11 points. The main conclusion is that our KGE-MMSLDA modeled the two modalities' information and knowledge prior information in a unified topic model. Meanwhile, KGE-MMSLDA performed better than Link-KGE-LDA and Corr-KGE-LDA, which means that incorporating supervised information into the topic model using the max-margin classifier and multi-modal information was helpful for mining interpretable topics.

Table II shows textual topics together with their PMI topic coherence scores mined from our collected dataset using the LDA and the KGE-MMSLDA models. We selected the five best-matched topics from the two models. For closely related topics, the topic words learned by the models were similar. The following conclusions can be drawn from the table as well, however: First, for the same topic, the PMI score calculated using our topic model was higher than that obtained by the traditional LDA. Second, our model mined unique words in the same topic. For the first topic—"Wall Street"—the KGE-MMSLDA found "square," which was not discovered by the traditional LDA. Similarly, for several other topics, such as "Maritime Dispute," "World Cup," "Politics," and "Military," the KGE-MMSLDA found "Bali," "Brasil," "peace," and "veterans," which were missed by the LDA model.

2) Event Classification Evaluation: To verify the performance of the KGE-MMSLDA, we used the libSVM classifier [59] to train event features. We compared the KGE-MMSLDA with several related baseline methods. For each, we set the symmetric Dirichlet priors to $\alpha = 1$,$\beta_w = 0.01$, and $\beta_v = 0.01$. We conducted a series of experiments with different numbers of topics and used the best classification accuracy of each method for comparison.

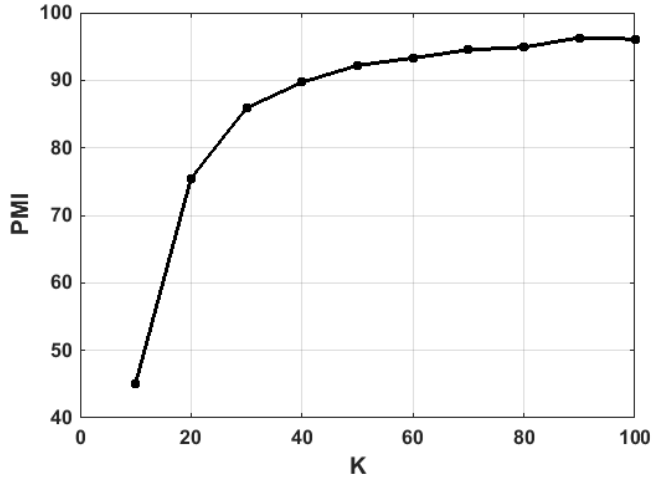


Fig. 4. PMI performance varying with different K values.

Social Event Classification: In the classification phase, the training and testing sets were setup using the following strategy: We split the dataset into five subsets and then iteratively ran all the methods on them using five-fold cross-validation. In each run of each algorithm, we iteratively selected one subset as the testing set and remaining four as training sets. Table III shows the average accuracy of the methods over the five subsets. To verify the performance of the KGE-MMSLDA for event classification, we added the results for a state-of-the-art supervised topic model (MMSLDA) to Table III for comparison. The LDA delivered the worst performance and the proposed KGE-MMSLDA the best of all methods. Link-KGE-LDA and Corr-KGE-LDA performed better than the traditional LDA, which means that using entity embedding learned from a large-scale knowledge base helped them implement discriminating event representation. The MMLDA outperformed the LDA because the modeling of text and images yield better results of feature representation than modeling text only. The KGE-MMSLDA outperformed the Corr-KGE-LDA, which means that supervised information was useful for social event classification. It also outperformed the MMSLDA and the other four methods, which shows that integrating a max-margin classifier and multi-modal feature representation into a unified topic model significantly helped improve classification performance, and knowledge embedding helped with event classification.

Number of Topics *K*: We explored the impact of the number of topics K on event classification performance. An appropriate number of topics K can help us determine a balance between the training time of the topic model and event classification performance. We evaluated KGE-MMSLDA's accuracy together with that of mainstream methods using different numbers of topics. Each model's accuracy is shown in Fig. 5, where the value of parameter K was changed from 10 to 50. As shown in the figure, our proposed KGE-MMSLDA model stably outperformed the other methods. Note that its classification accuracy increased rapidly when K was changed from 1 to 25, and remained stable when K s wan the range of 30–50. The recommended value for K is 40 for our dataset.

3) Multi-Modal Topic Mining: In this study, textual information and visual information were modeled using a unified KGE-MMSLDA. To show the multi-modal topic mining capability of the KGE-MMSLDA, we qualitatively show its mined topics in text and images in Fig. 6.

In a qualitative experiment, we set K to 40 and selected four topics for analysis and visualization. We first sorted the text topic words and image topic words by $p(w|z)$ and $p(v|z)$, respectively, and list the top six topics and their corresponding images for comparison. From Fig. 6, most learned textual words and images were interpretable, such as Adele, Syria, ISIS, NSA, and surveillance, and were clearly related to certain events. For example, the textual words in topic 10 were Adele,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

TABLE II
LEARNED TOPICS OF TRADITIONAL LDA AND KGE-MMSLDA MODEL.

LDA					KGE-MMSLDA				
street	sea	2014	korea	army	Day	sea	brazil	korea	army
occupy	tune	cup	china	service	street	island	2014	north	service
wall	rabbit	world	north	soldiers	occupy	bali	cup	syria	event
day	island	fifa	south	military	square	rabbit	world	war	military
protest	york	winner	asia	event	nepal	camel	brasil	korean	veterans
photos	hotels	football	korean	force	Wall	dutch	fifa	syrian	home
march	coney	soccer	director	families	temple	york	united	dprk	force
ows	dr	freestyle	chinese	memorial	protest	coney	states	civil	families
2011	hotel	life	institute	april	earthquake	yamada	estados	pyongyang	family
nyc	yamada	worldcup	japan	war	Photos	balinese	unidos	peace	lynch
109.7	61.23	58.1	91.4	90.2	113.7	64.5	70.5	155.9	75.1

TABLE III
CLASSIFICATION ACCURACY OF SOME MAINSTREAM METHOD.

Methods	Accuracy
LDA	0.682
MMLDA	0.725
Link-KGE-LDA	0.713
Corr-KGE-LDA	0.710
MMSLDA	0.763
KGE-MMSLDA	**0.851**

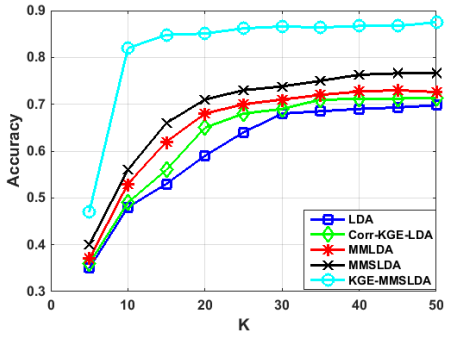


Fig. 5. Accuracy comparison of different methods with different topic number.

Fig. 6. Mined topics of KGE-MMSLDA.

live, concert, tour, love, and blue. Clearly, this topic was very relevant to the event "Adele concert tour." This relevance was demonstrated by the images in Topic 10. Similarly, topics 3, 7, and 33 were closely associated with the Syrian Civil War, PRISM surveillance program, and the Afghanistan War according to Fig. 6. To summarize, the KGE-MMSLDA can effectively mine multi-modal topics. It mined interpretable textual topics and described specific events through textual words and visual patches.

V. CONCLUSIONS

In this paper, We proposed a knowledge embedding-based max-margin topic model for multi-modal event analysis called the KGE-MMSLDA that combines knowledge embedding with the multi-modal topic model to jointly learn an event's representation and the classifier in a unified framework. The KGE-MMSLDA not only exploits multi-media data with prior knowledge to mine more coherent and meaningful topics with good representation, but also integrates supervised information into the multi-modal topic model as a regularization term to obtain discriminative representation for social event classification. The results of experiments on a large multi-modal dataset demonstrated that the KGE-MMSLDA outperforms mainstream approaches for social event analysis and classification. Despite its effectiveness, the time complexity of the proposed method is high, because it requires three sampling processes for text, image, and knowledge. Two possible improvements should be explored in future work: (1) optimizing the sampling strategy to boost sampling speed; and (2) determining other kind of information to incorporate into the model to improve accuracy.

ACKNOWLEDGMENT

This work is supported by the National Key Research and Development Program of China (No. 2017YFB0803301) and the National Natural Science Foundation of China (No. 61772170). The authors would like to thank the anonymous reviewers for their reviewing efforts and valuable comments.

REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," pp. 37–45, 1998.
[2] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 88–101, 2012.
[3] T. Zhang and C. Xu, "Cross-domain multi-event tracking via co-pmht," *Acm Transactions on Multimedia Computing Communications & Applications*, vol. 10, no. 4, pp. 1–19, 2014.
[4] X. Yang, T. Zhang, C. Xu, and M. S. Hossain, "Automatic visual concept learning for social event understanding," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 346–358, 2015.
[5] D. Patel, W. Hsu, and M. L. Lee, "Mining relationships among interval-based events for classification," in *ACM SIGMOD International Conference on Management of Data*, 2008, pp. 393–404.
[6] X. Wu, C. W. Ngo, and A. G. Hauptmann, "Multimodal news story clustering with pairwise visual near-duplicate constraint," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 188–199, 2008.

[7] I. Kalamaras, A. Drosou, and D. Tzovaras, "Multi-objective optimization for multimodal visualization," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1460–1472, 2014.

[8] W. Meng, L. Hao, T. Dacheng, L. Ke, and W. Xindong, "Multimodal graph-based reranking for web image search," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 21, no. 11, p. 4649, 2012.

[9] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, "Topic-conditioned novelty detection," in *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 688–693.

[10] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Information Retrieval*, vol. 7, no. 3-4, pp. 347–368, 2004.

[11] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Visual Analytics Science and Technology*, 2010, pp. 115–122.

[12] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 127–134.

[13] H. Gao, S. Tang, Y. Zhang, D. Jiang, F. Wu, and Y. Zhuang, "Supervised cross-collection topic modeling," in *ACM Multimedia*, 2012, pp. 957–960.

[14] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Spatial-disclda for visual recognition," in *Computer Vision and Pattern Recognition*, 2011, pp. 1769–1776.

[15] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora," in *Conference on Empirical Methods in Natural Language Processing: Volume*, 2009, pp. 248–256.

[16] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification." *Proceedings of NIPS Neural Information Processing Systems (2008*, pp. 897–904, 2008.

[17] Y. Bao, N. Collier, and A. Datta, "A partially supervised cross-collection topic model for cross-domain text classification," in *ACM International Conference on Information & Knowledge Management*, 2013, pp. 239–248.

[18] D. M. Blei and J. D. Mcauliffe, "Supervised topic models," *Advances in Neural Information Processing Systems*, vol. 3, pp. 327–332, 2010.

[19] Y. Wang and G. Mori, "Max-margin latent dirichlet allocation for image classification and annotation," *Lecture Notes in Computer Science*, vol. 1674, no. 1, pp. 39–48, 2011.

[20] D. Andrzejewski and X. Zhu, "Latent dirichlet allocation with topic-in-set knowledge," in *NAACL Hlt 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, 2009, pp. 43–48.

[21] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht, "A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic," in *International Joint Conference on Artificial Intelligence*, 2011, pp. 1171–1177.

[22] L. Yao, Y. Zhang, B. Wei, H. Qian, and Y. Wang, "Incorporating probabilistic knowledge into topic models," in *PAKDD*, 2015, pp. 586–597.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, Mar. 2003.

[24] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 297–304.

[25] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE transactions on multimedia*, vol. 18, no. 2, pp. 233–246, 2016.

[26] Y. Gao, H. Zhang, X. Zhao, and S. Yan, "Event classification in microblogs via social tracking," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 3, p. 35, 2017.

[27] D. Lin and J. Xiao, "Characterizing layouts of outdoor scenes using spatial topic processes," pp. 841–848, 2013.

[28] F. Rodrigues, M. Lourenco, B. Ribeiro, and F. C. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2409–2422, 2017.

[29] J. Zhu, N. Chen, H. Perkins, and B. Zhang, "Gibbs max-margin topic models with data augmentation." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1073–1110, 2014.

[30] S. Yang, C. Yuan, B. Wu, W. Hu, and F. Wang, "Multi-feature max-margin hierarchical bayesian model for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1610–1618.

[31] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent dirichlet allocation," *Advances in Neural Information Processing Systems*, vol. 23, pp. 856–864, 2010.

[32] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent dirichlet allocation for tag recommendation," in *ACM Conference on Recommender Systems, Recsys 2009, New York, Ny, Usa, October*, 2009, pp. 61–68.

[33] A. Perotte, N. Bartlett, N. Elhadad, and F. Wood, "Hierarchically supervised latent dirichlet allocation," *Advances in Neural Information Processing Systems*, vol. 24, pp. 2609–2617, 2011.

[34] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *Computer Vision and Pattern Recognition*, 2010, pp. 3408–3415.

[35] J. Sang and C. Xu, "Right buddy makes the difference: an early exploration of social relation analysis in multimedia applications," in *ACM International Conference on Multimedia*, 2012, pp. 19–28.

[36] K. S. Prabhudesai, B. O. Mainsah, L. M. Collins, and C. S. Throckmorton, "Augmented latent dirichlet allocation (lda) topic model with gaussian mixture topics," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2451–2455.

[37] S. Qian, T. Zhang, and C. Xu, "Multi-modal supervised latent dirichlet allocation for event classification in social media," 2014, pp. 152–157.

[38] L. Yao, Y. Zhang, B. Wei, Z. Jin, R. Zhang, Y. Zhang, and Q. Chen, "Incorporating knowledge graph embeddings into topic modeling," in *AAAI Conference on Artificial Intelligence*, 2017.

[39] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," in *AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, Usa, August*, 2011.

[40] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in Neural Information Processing Systems*, pp. 2787–2795, 2013.

[41] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014, pp. 1112–1119.

[42] S. Guo, Q. Wang, B. Wang, L. Wang, and L. Guo, "Semantically smooth knowledge graph embedding," in *Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015, pp. 84–94.

[43] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions." in *AAAI*, 2016, pp. 2659–2665.

[44] Z. Chen and B. Liu, "Mining topics in documents:standing on the shoulders of big data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1116–1125.

[45] F. Doshi-Velez, B. C. Wallace, and R. Adams, "Graph-sparse lda: a topic model with structured sparsity," *Computer Science*, 2014.

[46] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," vol. 3, pp. 299–313, 2015.

[47] R. Das, M. Zaheer, and C. Dyer, "Gaussian lda for topic models with word embeddings," in *Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015, pp. 795–804.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[53] K. Batmanghelich, A. Saeedi, K. Narasimhan, and S. Gershman, "Nonparametric spherical topic modeling with word embeddings," 2016.

[54] T. L. Griffiths and M. Steyvers, "Finding scientific topics," vol. 101, 2004, pp. 5228–5235.

[55] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, no. 2, pp. 1107–1135, 2003.

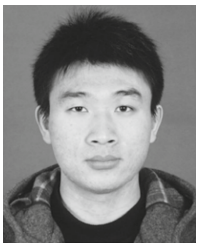
[56] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539.
[57] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. Mccallum, "Optimizing semantic coherence in topic models." in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 262–272.
[58] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, 2010, pp. 100–108.
[59] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," vol. 2, no. 3, pp. 1–27, 2011.

Feng Xue Dr. Feng Xue is a professor with the Hefei University of Technology(HFUT). He received his Ph.D. degree(June 2006) from the Dept. of Computer Science of Hefei University of Technology. His current research interests are in artificial intelligence, multimedia analysis and recommendation system.

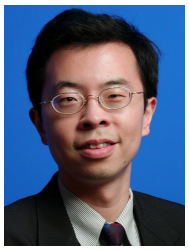
Richang Hong Richang Hong (M'12) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He is currently a Professor with the Hefei University of Technology, Hefei, China. He was a Research Fellow with the School of Computing, National University of Singapore from 2008 to 2010. He has co-authored more than 100 publications in the areas of his research interests, which include multimedia content analysis and social media. He was a recipient of the Best Paper Award in the ACM Multimedia 2010, Best Paper Award in the ACM ICMR 2015, and the Honorable Mention of the IEEE Transactions on Multimedia Best Paper Award. He served as the Associate Editor of the IEEE Multimedia Magazine, Information Sciences and Signal Processing, Elsevier and the Technical Program Chair of the MMM 2016. He is a member of ACM and the Executive Committee Member of the ACM SIGMM China Chapter.

Xiangnan He Dr. Xiangnan He is a professor with the University of Science and Technology of China (USTC). He received his Ph.D. in Computer Science from National University of Singapore (NUS) in 2016. His research interests span information retrieval, data mining, and multi-media analytics. He has over 60 publications appeared in several top conferences such as SIGIR, WWW, and MM, and journals including TKDE, TOIS, and TMM. His work on recommender systems has received the Best Paper Award Honourable Mention in WWW 2018 and ACM SIGIR 2016. Moreover, he has served as the PC chair of CCIS 2019, area chair of MM 2019 and CIKM 2019, and PC member for several top conferences including SIGIR, WWW, KDD etc., and the regular reviewer for journals including TKDE, TOIS, TMM, etc.

Jianwei Wang Jianwei Wang is now a big data development engineer. His research interest lies in multimedia analysis and data governance. He received the master degree from the Computer Science of Hefei University of Technology in 2018.



Shengsheng Qian Shengsheng Qian received the B.E. degree from the Jilin University, Changchun, China, in 2012, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include social media data mining and social event content analysis.



Changsheng Xu Changsheng Xu (M'97–SM'99–F'14) is a Professor in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and Executive Director of China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. He has hold 30 granted/pending patents and published over 200 refereed research papers in these areas. Dr. Xu is an Associate Editor of IEEE Trans. on Multimedia, ACM Trans. on Multimedia Computing, Communications and Applications and ACM/Springer Multimedia Systems Journal. He received the Best Associate Editor Award of ACM Trans. on Multimedia Computing, Communications and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He served as Program Chair of ACM Multimedia 2009. He has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops. He is IEEE Fellow, IAPR Fellow and ACM Distinguished Scientist.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

VI. Appendix: Inference for Conditional Distribution of Textual, Visual and Knowledge Embedding Topic Assignment Variable Z_d^w, Z_d^v

According to the (4) in this paper, the normalized posterior distribution of KGE-MMSLDA can be formulated as:

$$q(\eta,\lambda,Z,\Theta,\Phi_w,\Phi_v) = \frac{p_0(\eta,\lambda,\Theta,\Phi_w)\cdot p(w,v,e \mid Z,\Phi_w,\Phi_v)\varphi(y,\lambda \mid Z,\eta)}{\Psi(y,w,v,e)} \quad (19)$$

Based on the data augment formula of Gibbs MedLDA in [29], by integrating out the Dirichlet variables(Θ,Φ_w,Φ_v), (19) can be simplified as the collapsed posterior distribution as (20).

$$\begin{aligned} q(\eta,\lambda,Z) &\propto p_0(\eta)\cdot p(w \mid Z)\cdot\prod_{i=1}^{L}\varphi_i(y_i,\lambda_i|Z,\eta) \\ &= p_0(\eta)\cdot\frac{p(Z \mid w)\cdot p(w)}{p(Z)}\cdot\prod_{i=1}^{L}\varphi_i(y_i,\lambda_i \mid Z,\eta) \\ &\propto p_0(\eta)\cdot p(Z|w)\cdot\prod_{i=1}^{L}\varphi_i(y_i,\lambda_i \mid Z,\eta) \end{aligned} \quad (20)$$

where w,v and e represent textual, visual and knowledge embedding input, respectively; η denotes the weights of classifier, λ denotes the augmented variable, y_i is a real label, L is the number of classifiers. As $p_0(\eta)$ is a known prior distribution, we ignore this term and (20) can be written as:

$$\begin{aligned} p(Z_d^w = k \mid Z_{\neg(d)}^w, w,v,e,\eta,\lambda) &\propto p(Z_d^w = k \mid Z_{\neg(d)}^w, w)\cdot\prod_{i=1}^{L}\varphi_i(y_i,\lambda_i \mid Z,\eta) \end{aligned} \quad (21)$$

where $p(Z_d^w = k \mid Z_{\neg(d)}^w, w,v,e,\eta,\lambda)$ is the product of two parts: LDA model and Gibbs regression model. In the following, we will derive the two parts respectively. As image input v and knowledge entity input e are irrelated to the conditional distribution of Z_d^w, the first part in the right of (21) can be rewritten as

$$\begin{aligned} p(Z_d^w = k \mid Z_{\neg(d)}^w, w) &= \frac{p(w \mid z)}{p(w_{\neg(d)} \mid z_{\neg(d)})\cdot p(w_d)}\cdot\frac{p(z)}{p(z_{\neg(d)})} \\ &\propto \frac{\delta(\eta_z+\beta_w)}{\delta(n_{z,\neg(d)}+\beta_w)}\cdot\frac{\delta(\eta_k+\alpha)}{\delta(n_{k,\neg(d)}+\alpha)} \\ &\propto \frac{\Gamma(n_{k,t}+\beta_w)\cdot\Gamma(\sum_{t=1}^{D^w}(n_{k,t}^{\neg(w)}+\beta_w))}{\Gamma(n_{k,t}^{\neg(w)}+\beta_w)\cdot\Gamma(\sum_{t=1}^{D^w}(n_{k,t}+\beta_w))}\cdot \\ &\quad \frac{\Gamma(n_{d,k}^w+\alpha)\cdot\Gamma(\sum_{k=1}^{K}(n_{d,k}^{\neg(w)}+\alpha))}{\Gamma(n_{d,k}^{\neg(w)}+\alpha)\cdot\Gamma(\sum_{k=1}^{K}(n_{d,k}^w+\alpha))} \\ &\propto \frac{n_{k,t}^{\neg(w)}+\beta_w}{\sum_{t=1}^{D_w}(n_{k,t}+\beta_w)}\cdot\frac{n_{d,k}^{\neg(w)}+\alpha}{\sum_{k=1}^{K}(n_{d,k}^w+\alpha)} \end{aligned} \quad (22)$$

where

$$\delta(x) = \frac{\prod_{i=1}^{dim(x)}\Gamma(x_i)}{\Gamma(\prod_{i=1}^{dim(x)}x_i)}$$

$\Gamma(\cdot)$ is the Gamma function, $n_{d,k}^{\neg(w)}$ is the number of textual words assigned to the latent topic k of document d, except current textual word w, $n_{d,k}^w$ is the number of textual words assigned to the latent topic k of document d, $n_{k,t}^{\neg(w)}$ is the number of textual word t of topic k that appear in all documents, except the current textual word w. $n_{k,t}$ is the number of textual words t of topic k that appear in all documents; α and β_w are the priors of the Dirichlet distribution.

In fact, the second part of the right in (21) can be seen as a cumulative product of a series of classifiers since our model is used for L classification tasks.

As defined in the Lemma 2(Scale Mixture Representation) of [29], the individual distribution of a single classifier can be expressed as

$$\begin{aligned} \varphi_i(y_i,\lambda_i \mid Z,\eta) &= \prod_{d=1}^{D}\frac{1}{\sqrt{2\pi\lambda_d^i}}exp(-\frac{(\lambda_d^i+c\zeta_d^i)^2}{2\lambda_d^i}) \\ &= N(\eta_i;\mu_i,\varepsilon_i) \end{aligned} \quad (23)$$

where $\zeta_d^i = l - y_d^i\eta_{i,k}$, $\overline{z_d}, l$ denotes the cost of making a wrong prediction, y^i is a real label, $\eta_{i,k}$ is the weight of the classifier, λ denotes the augmented variable, $\overline{z_d} = \frac{1}{N}\sum_{n=1}^{N}Z_{n,d}$ denotes average topic assignment of document d. ε_i is the posterior covariance matrix and μ_i is the posterior mean, which can be calculated by using following formulation. Readers are referred to the Section 4 of [29]) for details:

$$\varepsilon_i = (\frac{1}{v^2}+c^2\sum_{d=1}^{D}\frac{\overline{z_d}\cdot\overline{z_d}^T}{\lambda_d^i})^{-1} \quad (24)$$

$$\mu_i = \varepsilon_i(c\sum_{d=1}^{D}y_d^i\frac{\lambda_d^i+c\mathfrak{l}}{\lambda_d^i}\overline{z_d}) \quad (25)$$

According to (23), (24) and (25), Gibbs regression model can be expressed as

$$\varphi_i(y_i,\lambda_i \mid Z,\eta) = exp(\frac{y_d^i c(cT+\lambda_d^i)\eta_{i,k} - \frac{2N_d^w-2}{N_d^w}\Lambda_{d,n}^i\eta_{i,k} - \frac{c^2}{2N_d^w}\eta_{i,k}^2}{N_d^w\lambda_d^i}) \quad (26)$$

where $\Lambda_{d,n}^i\eta_{i,k} = \frac{1}{N_d-1}\sum_{k=1}^{K}\eta_{i,k}n_{d,\neg n}$ denotes the discriminant function value of the words except the current word n, N_d^w is the number of textual words for document d, η denotes the classifier weights, λ denotes the augmented variable, y_i is a real label.

Thus, by combining the derivation formulas of the above two parts, we can get the conditional distribution of Z_d^w.

$$\begin{aligned} p(Z_d^w = k \mid Z_{\neg(d)}^w, w,v,e,\eta,\lambda) &\propto \frac{n_{k,t}^{\neg(w)}+\beta_w}{\sum_{t=1}^{D_w}(n_{k,t}+\beta_w)}\cdot\frac{n_{d,k}^{\neg(w)}+\alpha}{\sum_{k=1}^{K}(n_{d,k}^w+\alpha)} \\ &\cdot\prod_{i=1}^{L}exp(\frac{y_d^i c(cT+\lambda_d^i)\eta_{i,k} - \frac{2N_d^w-2}{N_d^w}\Lambda_{d,n}^i\eta_{i,k} - \frac{c^2}{2N_d^w}\eta_{i,k}^2}{N_d^w\lambda_d^i}) \end{aligned} \quad (27)$$

In the same way, the inference of conditional distribution of visual topic assignment variable Z_d^v is similar to that of Z_d^w,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2951194, IEEE Transactions on Multimedia

which is a product of LDA model and Gibbs regression model according to (20).

$$p(Z_d^v = k \mid Z_{\neg(d)}^v, w, v, e, \eta, \lambda) \propto p(Z_d^v = k \mid Z_{\neg(d)}^v, v) \cdot \prod_{i=1}^{L} \varphi_i(y_i, \lambda_i \mid Z, \eta) \tag{28}$$

Similar to (22) and (23), $p(Z_d^v = k \mid Z_{\neg(d)}^v, v)$ and $\varphi_i(y_i, \lambda_i \mid Z, \eta)$ can be defined as

$$p(Z_d^v = k \mid Z_{\neg(d)}^v, v) \propto \frac{n_{k,t}^{\neg(v)} + \beta_v}{\sum_{t=1}^{D_v}(n_{k,t} + \beta_v)} \cdot \frac{n_{d,k}^{\neg(v)} + \alpha}{\sum_{k=1}^{K}(n_{d,k}^v + \alpha)} \tag{29}$$

$$\varphi_i(y_i, \lambda_i \mid Z, \eta) = exp\left(\frac{y_d^i c(cT + \lambda_d^i)\eta_{i,k} - \frac{2N_d^v - 2}{N_d^v}\Lambda_{d,n}^i \eta_{i,k} - \frac{c^2}{2N_d^v}\eta_{i,k}^2}{N_d^v \lambda_d^i}\right) \tag{30}$$

where N_d^v is the number of visual words for document d. η denotes the classifier weights, λ denotes the augmented variable, y_i is a real label. $n_{d,k}^{\neg(v)}$ is the number of visual words assigned to the latent topic k of document d, except current visual word v, $n_{d,k}^v$ is the number of visual words assigned to latent topic k of document d; $n_{k,t}^{\neg(v)}$ is the number of visual word t of topic k that appear in all documents, except the current visual word v; $n_{k,t}$ is the number of visual words t of topic k that appear in all documents; and α and β_v are the priors of the Dirichlet distribution.

We can derive the conditional distribution equation of Z_d^v by combining (29) and (30).

$$p(Z_d^v = k \mid Z_{\neg(d)}^v, w, v, e, \eta, \lambda) \propto \frac{n_{k,t}^{\neg(v)} + \beta_v}{\sum_{t=1}^{D_v}(n_{k,t} + \beta_v)} \cdot \frac{n_{d,k}^{\neg(v)} + \alpha}{\sum_{k=1}^{K}(n_{d,k}^v + \alpha)} \cdot \prod_{i=1}^{L} exp\left(\frac{y_d^i c(cT + \lambda_d^i)\eta_{i,k} - \frac{2N_d^v - 2}{N_d^v}\Lambda_{d,n}^i \eta_{i,k} - \frac{c^2}{2N_d^v}\eta_{i,k}^2}{N_d^v \lambda_d^i}\right) \tag{31}$$