Cross-media Topic Detection with Refined CNN based Image-Dominant Topic Model

Zhiyi Wang[†♮], Liang Li[†♮], Qingming Huang[†‡♮]
[†]University of Chinese Academy of Sciences, China
[‡]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, China
[♮]Key Laboratory of Big Data Mining and Knowledge Management, CAS, China
{zhiyi.wang, liang.li}@vipl.ict.ac.cn qmhuang@ucas.ac.cn

ABSTRACT

Online heterogenous data is springing up while the data has the rich auxiliary information (e.g. pictures and videos) around the text. However, traditional topic models are suffering from the limitations to discover the topics effectively from the cross-media data. Incorporating with the convolutional neural network (CNN) feature, we propose a novel image dominant topic model, which projects both the text modality and the visual modality into a semantic simplex. Further, an improved CNN feature is introduced to capture more visual details by fusing the convolutional layer and fully-connected layer. Experimental comparisons with state-of-the-art methods in the cross-media topic detection task show the effectiveness of our model.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Experimentation, Theory

Keywords

Convolutional Neural Networks, Topic Model, Cross-media topic Detection

1. INTRODUCTION

Recently, social networks, video and content sharing websites burst to produce a large number of online multimedia data every day, and more and more people tend to share their moments with the heterogeneous information (text, image and video). In such expression of cross-media, the visual data usually delivers the more entitative content while the text data usually describes the emotion. Thus, how to use cross-media data properly and effectively is the key to detect the topics in the real networks.

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806309.

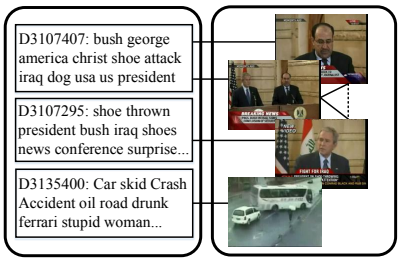

Figure 1: An illustration of semantic relationship in document. Explicit and potential relationships are drawn by solid and dotted lines respectively.

The visual component of cross-media data plays an important role in providing the rich and useful information for the topic detection. Deep neural networks [8, 15] are the most popular models for the effective representation of the visual information. Since large convolutional network models have demonstrated impressive classification performance on the ImageNet benchmark [8], deep neural networks become newborn and popular methods to learn features, especially in the visual domain.

The topic model is widely used to solve the semantic-related tasks. Probabilistic topic models [2, 6] have been powerful methods of learning topic representation of document corpus. To treat the multi-modality information, [12] models the joint distribution of image content and class labels with supervised LDA. However, these topic models make the strong assumption of independent identical distribution of documents while they also ignore the hidden association among the cross-media data. To further solve this problem, Author Topic Model (ATM) [11] and Relational Topic Model (RTM) [4] are proposed to model the content of documents with link relationships. To model the cross-media data, Semi-supervised Relational Topic Model (ss-RTM) [10] incorporates image content, labels and relations into topic modeling. Hashtag Graph based Topic Model for Tweet Mining (HGTM) [13] introduces a hashtag relation graph as weakly-supervised information for effective tweet semantic modeling.

Although these methods have achieved some performance improvements, most previous works exploit document-oriented relationships to formulate the topic model. A beneficial fact is that more and more people express themselves by combining images and a few text in the social network and cross-media platforms. Obviously, straightforward and information-

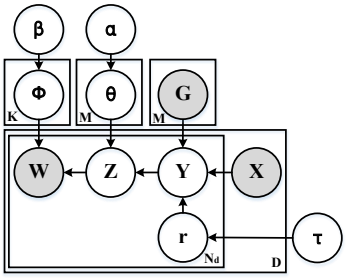

Figure 2: The graphical representation of IDTM.

rich images can reveal more relationships than the text. As shown in Figure 1, the images are weakly annotated in the social network, but the images of the same topic can enhance the latent relationships among the short and noisy documents.

Inspired by the above insights, incorporating with the deep feature of the visual information, this paper proposes an Image-Dominant Topic Model (IDTM), where the text is projected into semantic space by the latent features via the corresponding images. IDTM is a generative probabilistic model that incorporates weakly-supervised information based on a weighted image graph. By establishing the weighted image graph, IDTM can discover explicit text-image relationships and potential text-image relationships. Furthermore, we introduce a fuse mechanism of convolutional Layer 5 and fully-connected Layers 7 from the CNN feature based on the framework of Caffe [7], which is to capture more visual details and improve the generalization of the final visual representation. Finally, we evaluate the proposed method in the cross-media topic detection task and compare with several state-of-the-art methods to show its effectiveness.

The remainder of the paper is organized as follows. Section 2 describes IDTM for topic understanding and coarse-to-fine method. Section 3 analyzes the experimental results. Finally, Section 4 concludes this paper.

2. METHODOLOGY

The model for IDTM is shown in Figure 2. We define the number of documents and topics in a corpus as D and K. Suppose the document d corresponds to a word index $\mathbf{w}_d = \{w_{d1}, w_{d2}, \ldots, w_{dN_d}\}$, where N_d is the number of words in document d, and an image index $\mathbf{x}_d = \{x_{d1}, x_{d2}, \ldots, x_{dM_d}\}$, where M_d is the number of images in document d. We allocate a topic assignment z_{di} and an image assignment y_{di} for each w_{di} in the document d. Each topic by a Multinomial distribution over words as ϕ with β as the Dirichlet hyper-parameter. Similarly, each image by a Multinomial distribution over topics as θ with α as the Dirichlet hyper-parameter. τ is a Bernoulli variable to decide whether to assign images in the document d for current w_{di}. An image relation graph is an undirected graph, denoted as $G = (V, E)$ where nodes $V = \{m\}_{m=1}^{M}$ are images, and each edge e_{ij} of the similarity matrix $E = \{(e_{ij})\}_{i,j \in V, i \neq j}$ means the similarity between image i and image j. The shaded nodes indicate observations, while the others represent latent variables. The details of the generative process are as follows:

1. For each topic k: $\phi_k \sim Dirichlet(\beta)$
2. For each image x: $\theta_x \sim Dirichlet(\alpha)$
3. For each document $d \in D$

   For each word w_{di}, $i = 1 : N_d$

   a. Sample an image $y_{di}^1 \sim Uniform(\mathbf{x}_d)$

   b. Sample $r \sim Bernoulli(\tau)$

   c. If $r = 1$,

   sample $y_{di} = y_{di}^1$

   If $r = 0$,

   sample $y_{di} \sim Multinomial(norm(g_{y_{di}^1}))$

   d. Sample a topic $z_{di} \sim Multinomial(\theta_{y_{di}})$

   e. Sample a word $w_{di} \sim Multinomial(\phi_{z_{di}})$

$g_{y_{di}^1}$ denotes the image relation graph and is normalized by column. Hence, $norm(g_{y_{di}^1})$ represents the compactness of relationships between images and is an association probability vector, where the j^{th} element is:

$$p(y_j | y_{di}^1) = \frac{g_{y_{di}^1, y_j}}{\sum_{j'} g_{y_{di}^1, y_{j'}}} \tag{1}$$

In the generative process, the key step is to sample a related image for the current word. Thus, we model the process by two steps: **1)** an image y_{di}^1 is sampled uniformly from $\mathbf{x}_d$; **2)** the parameter r is sampled from Bernoulli distribution to determine whether the current word is related to images $\mathbf{x}_d$. If related ($r = 1$), y_{di} equals to y_{di}^1; if not ($r = 0$), we sample y_{di} from the multinomial distribution of $norm(g_{y_{di}^1})$.

2.1 Inference and Parameter Estimation

In this section, we come to the learning strategy of our model. Words, images and image relationships are observed while the hidden variables are guided by latent distribution parameters. With the model defined, we turn to evaluate the posterior inference and parameter estimation. The probability of the whole corpus is:

$$p(\mathbf{w} | \theta, \phi, \mathbf{r}, \mathbf{x}, \mathbf{G}) = \prod_{d=1}^{D} p(\mathbf{w}_d | \theta, \phi, \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \tag{2}$$

Assuming that "topic-word" distribution and "image-topic" distribution are conditionally independent, we have:

$$\begin{aligned} p(\mathbf{w}_d | \theta, \phi, \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) &= \prod_{i=1}^{N_d} p(w_{di} | \theta, \phi, \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \\ &= \prod_{i=1}^{N_d} \sum_{s=1}^{M_d} \sum_{k=1}^{K} p(w_{di}, z_{di} = k, y_{di} = s | \theta, \phi, \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \\ &= \prod_{i=1}^{N_d} \sum_{s=1}^{M_d} \sum_{k=1}^{K} p(w_{di} | z_{di} = k, \phi) p(z_{di} = k | y_{di} = s, \theta) \cdot \\ &\qquad p(y_{di} = s | \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \\ &= \prod_{i=1}^{N_d} \sum_{s=1}^{M_d} \sum_{k=1}^{K} \phi_{w_{di} k} \theta_{ks} p(y_{di} = s | \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \end{aligned} \tag{3}$$

$p(y_{di} = s | \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d})$ indicates the probability of image assignment s conditioned on the current explicit images $\mathbf{x}_d$ and related potential images. Based on the Bernoulli distribution that we describe above, the probability that we assign images to words is:

$$\begin{aligned} &p(y_{di} = s | \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \\ &= \left[p(y_{di}^1 = s | \mathbf{x}_d) p(y_{di} = s | y_{di}^1) \right]^r \cdot \\ &\left[\sum_{j=1}^{M_d} p(y_{di}^1 = x_{d,j} | \mathbf{x}_d) p(y_{di} = s | y_{di}^1 = x_{d,j}, g_{x_{d,j}}) \right]^{1-r} \end{aligned} \tag{4}$$

$x_{d,j}$ is the j^{th} image in document d. As the exact inference of the model is general intractable, some commonly used approximate methods are usually conducted in parameter inference as substitutes, such as variational inference[1], expectation propagation [9], or Gibbs sampling [5]. In this paper, we employ Gibbs sampling procedure to carry out approximated parameters for its simplicity and effectiveness and obtain the sample posterior distribution:

$$p(z_{di} = k, y_{di} = s, r_{di} = u | w_{di} = w, \mathbf{z}_{-di}, \mathbf{y}_{-di}, \mathbf{w}_{-di}, \mathbf{X}, \mathbf{G}, \alpha, \beta, \tau) \propto \frac{n_{k,w_{di}}^{-di} + \beta}{\sum_W n_{k,w}^{-di} + W\beta} \cdot \frac{n_{k,x_{di}}^{-di} + \alpha}{\sum_K n_{k,x}^{-di} + K\alpha} \cdot p(y_{di} = s | \mathbf{r}, \mathbf{x}_d, \mathbf{g}_{x_d}) \quad (5)$$

$n_{k,w_{di}}^{-di}$ is the occurrence number of word w_{di} assigned to topic k except for the current word while $n_{k,x_{di}}^{-di}$ is the occurrence number of topic assigned to image x_{di} except for the current word. W and K are the number of words and topics in whole corpus respectively. After iterative sampling, the probability reaches the convergence. ϕ and θ are:

$$\phi_k \propto \frac{n_{k,w_{di}} + \beta}{\sum_W n_{k,w} + W\beta}, \qquad \theta_s \propto \frac{n_{k,x_{di}} + \alpha}{\sum_K n_{k,x} + K\alpha} \quad (6)$$

According to the topic structure detected, IDTM can conclude distinguishable topics in real-world datasets and find out the representative words for each topic. Meanwhile, we can obtain the related images under each topic.

2.2 Coarse-to-fine Topic Detection

To capture more visual details, we employ deep neural networks in this paper. Based on the results shown in [14], the feature hierarchies become deeper, they learn increasingly powerful features. Thus, higher layers generally produce more discriminative and fine features. With the analysis of [14], Layer 5 and Layer 7 in Caffe [7] both have a steady improvement with SVM or Softmax. Ulteriorly, we explore how coarse or fine the features in Layer 5 and Layer 7 are and how to improve the performance by refining Layer 7. As far as we know, the features of Layer 7 are computed as follow:

$$v_1 = W_6 \cdot CNN_5 + B_6, \qquad v_2 = W_7 \cdot v_1 + B_7 \quad (7)$$

where W_i is the weights and B_i is the biases of each layer learnt by [7]. v_1 and v_2 are the features of Layer 6 and Layer 7 respectively.

Assuming Layer 5 can be regarded as coarse features and omits some detailed information, albeit with multi-iteration. We propose a coarse-to-fine method to optimize these features. From the weights of each layer, some absolute values of the weights are close to zero while their contiguous eight weights are not. Thus, it comes to us that the current weight may also contain information in spite of being set to zero after multi-iteration. Above all, we replace the current weight by the average of contiguous eight weights. Our weighted image relation graph is generated as Figure 3. Images are represented as a set of CNN features from Layer n and the weighted image relation graph will be built by calculating the similarities of features.

3. EXPERIMENTS AND ANALYSIS

3.1 Experiment Setting

In this section, we evaluate our method on a cross-media dataset, MCG-WEBV [3]. MCG-WEBV is built with the "Most viewed" videos of "This month" on YouTube from

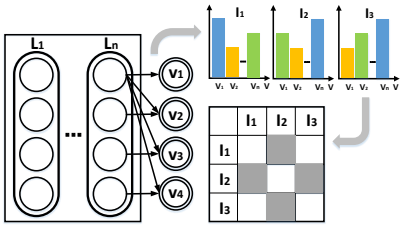

Figure 3: An illustration of generating weighted image relation graph.

Dec 2008 to Feb 2009. It contains 3282 web videos and 73 manually annotated ground truth topics. To compare our results with ground truth, we select the documents labeled from MCG-WEBV. Since the vocabulary has a very long tail word distribution, we filter out those words that occur less than or equal to 5 times. All text content is stemmed by portStemmer. Eventually, there are 832 documents, 22755 images and 4741 words in the collection.

In experiments, some preprocessing for establishing image relation graph is required. Firstly, to demonstrate the performance of the proposed method, we calculate cosine similarities on three kinds of features as comparisons: **1)** Bag of Visual Word features (abbreviated as BoVW). Each image is represented as a histogram of 128-codeword SIFT codebook; **2)** Original CNN features. We mainly analyze Layer 5 and Layer 7 of [7] (denote as CNN-5 and CNN-7); **3)** Coarse-to-fine CNN features (denote as CNN-7*). Secondly, we find out those images that have intense relationships (e.g. the value of cosine similarity ≥ 0.8) and set the other similarities to 0 to make the weighted image relation graph sparse. While preprocessing is relatively simple to deal with the relationships between images, it can reflect the potential relationships with different weights to some extent. We compare IDTM with two other models: **1)** LDA, which takes each text content as document; **2)** ATM, which treats images as "authors". In experiments, $\alpha = 50/K$ and $\beta = 0.01$ and Gibbs sampling is run for 1800 iterations.

3.2 Analysis of Our Approach

Here we analyze the quality of topics discovered by our model. Table 1 show examples of 2 topics out of 73 from MCG-WEBV dataset learnt by IDTM. We use the highest probability word to represent the topic.

Table 1: An illustration of 2 topics out of 73

Bush		Warcraft	
word	*prob*	*word*	*prob*
bush	0.16680	warcraft	0.03369
shoe	0.15164	ad	0.02379
president	0.09983	wow	0.02280
iraq	0.06193	paladin	0.02180
george	0.05940	tv	0.02081
iraqi	0.05561	commercial	0.01982
journalist	0.03918	pub	0.01784
throw	0.03413	spot	0.01685
conference	0.03160	ozzy	0.01685
press	0.02276	osbourne	0.01586

According to the results, the top-10 words and images are highly related to the specific topic. The corresponding top-

Figure 4: Image predictions by IDTM on topic Bush and Warcraft. The maximum and minimum probability of images is upper-left and lower-right respectively for each topic.

ic descriptions in [3] are "**Bush was attacked by shoes during press conference in Iraq**" and "**a popular online gaming world of warcraft (WOW in short), including the operate guides and ads**" respectively. The results denote that the top-5 words predicted by IDTM have explained the key information about the topic. For example, from "bush", "shoe", "president", "iraq", and "george", we might guess that topic is about "President George Bush" and "Iraq". The other words in top-10 may explain the information of where and who. Meanwhile, the top-10 images predicted by IDTM show the potential image relationship. As illustrated in Figure 4, images in the same topic are relative and gathered together, while they are in the separate documents beforehand. To demonstrate the performance of the proposed method, we compare the proposed model with ***F-Measure***. Table 2 shows the *F-Measure* of top-10 detected topics for all the test methods on MCG-WEBV dataset. ILDA and TLDA use Bag of visual Words of images and Bag of Words of documents as attributes respectively.

Table 2: *F-Measure* on MCG-WEBV

Method	IDTM			ATM	TLDA	ILDA
	$\tau = 0.8$	$\tau = 0.7$	$\tau = 0.6$			
BoVW	0.5244	0.5237	0.5184	0.4660	0.4565	0.1091
CNN-5	0.5499	0.5269	0.5317			
CNN-7	0.5486	0.5537	0.5255			
CNN-7*	0.5397	**0.5616**	0.5395			

From Table 2, the *F-Measure* of IDTM is significantly improved than other models no matter what hyper-parameter τ[1] or features we choose. The features in the first column are denoted in Section 3.1. Based on the method described in Section 2.2, we update W_6 to refine Layer 7. Since improved CNN feature shows better performance than BoVW and original CNN features generally, it convinces us that refining CNN feature by coarse-to-fine method indeed increases the discriminative power of images. Accordingly, the idea that combines text and image features together and builds topic model is meaningful and valuable.

4. CONCLUSIONS

In this paper, we propose a new deep image dominant topic model for cross-media topic detection. The model projects the text and the visual information into a semantic simplex. Compared to the traditional single document-oriented topic models, IDTM captures semantic relations between words by establishing weighted image relation graph and discovers more distinguishable topics than previous models. We introduce a fuse model of the convolutional layer and fully-connected layer in the CNN framework, which can obtain more visual details and improve the generalization of the final visual representation. In the future, we would like to explore image relationships with more robust methods and model emotional information among topics by image graph.

[1]In IDTM, hyper-parameter τ defines the possibility that image assignments are from images in the current document. The lower τ is, the higher randomness is, and vice versa.

5. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, in part by National Natural Science Foundation of China: 61332016, 61402431, and 61303154, in part by Project Funded by China Postdoctoral Science Foundation.

6. REFERENCES

[1] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, pages 127–134, 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] J. Cao, Y. D. Zhang, Y. C. Song, Z. N. Chen, X. Zhang, and J. T. Li. Mcg-webv: A benchmark dataset for web video analysis. *Beijing: Institute of Computing Technology*, 10:324–334, 2009.

[4] J. Chang and D. M. Blei. Relational topic models for document networks. In *AISTATS*, pages 81–88, 2009.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. *NAS*, 101(suppl 1):5228–5235, 2004.

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

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[9] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, pages 362–369, 2001.

[10] Z. X. Niu, G. Hua, X. B. Gao, and Q. Tian. Semi-supervised relational topic model for weakly annotated image recognition in social media. In *IEEE CVPR*, pages 4233–4240, 2014.

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494, 2004.

[12] C. Wang, D. M. Blei, and F. F. Li. Simultaneous image classification and annotation. In *IEEE CVPR*, pages 1903–1910, 2009.

[13] Y. Wang, J. Liu, J. S. Qu, Y. L. Huang, J. M. Chen, and X. Feng. Hashtag graph based topic model for tweet mining. In *IEEE ICDM*, pages 1025–1030, 2014.

[14] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. 2014.

[15] Y. Zheng, Y. J. Zhang, and L. Hugo. A deep and autoregressive approach for topic modeling of multimodal data. *arXiv preprint:1409.3970*, 2014.