

Bilateral Correspondence Model for Words-and-Pictures Association in Multimedia-rich Microblogs

ZHIYU WANG, Tsinghua University, Department of Computer Science and Technology
 PENG CUI, Tsinghua University, Department of Computer Science and Technology
 LEXING XIE, Australian National University and NICTA
 WENWU ZHU, Tsinghua University, Department of Computer Science and Technology
 YONG RUI, Microsoft Research Asia
 SHIQIANG YANG, Tsinghua University, Department of Computer Science and Technology

Nowadays, the amount of multimedia contents in microblogs is growing significantly. More than 20% of microblogs link to a picture or video in certain large systems. The rich semantics in microblogs provide an opportunity to endow images with higher-level semantics beyond object labels. However, this arises new challenges for understanding the association between multimodal multimedia contents in multimedia-rich microblogs. Disobeying the fundamental assumptions of traditional annotation, tagging, and retrieval systems, pictures and words in multimedia-rich microblogs are loosely associated and a correspondence between pictures and words cannot be established. To address the aforementioned challenges, we present the first study analyzing and modeling the associations between multimodal contents in microblog streams, aiming to discover multimodal topics from microblogs by establishing correspondences between pictures and words in microblogs. We first use a data-driven approach to analyze the new characteristics of the words, pictures and their association types in microblogs. We then propose a novel generative model, called the Bilateral Correspondence Latent Dirichlet Allocation (BC-LDA) model. Our BC-LDA model can assign flexible associations between pictures and words, and is able to not only allow picture-word co-occurrence with bilateral directions, but also single modal association. This flexible association can best fit the data distribution, so that the model can discover various types of joint topics and generate pictures and words with the topics accordingly. We evaluate this model extensively on a large-scale real multimedia-rich microblogs dataset. We demonstrate the advantages of the proposed model in several application scenarios, including image tagging, text illustration and topic discovery. The experimental results demonstrate that our proposed model can significantly and consistently outperform traditional approaches.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Social media, topic models, image analysis

ACM Reference Format:

Zhiyu Wang, Peng Cui, Lexing Xie, Wenwu Zhu, Yong Rui, and Shiqiang Yang. 2014. Bilateral Correspondence Model for Words-and-Pictures Association in Multimedia-rich Microblogs. *ACM Trans. Multimedia Comput. Commun. Appl.* 2, 3, Article 1 (February 2014), 20 pages.

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

Authors' addresses: Zhiyu Wang, Peng Cui, Wenwu Zhu, Shiqiang Yang, FIT 1-304, Tsinghua University, Beijing, China; email: zy-wang08@mails.tsinghua.edu.cn, {cuip,wwzhu,yangshq}@tsinghua.edu.cn; Lexing Xie, Bldg 108, North Rd, Room N326, The Australian National University, Canberra, ACT 0200, Australia; email: lexing.xie@anu.edu.au; Yong Rui, Microsoft Research Asia, No. 5, Dan Ling Street, Haidian District, Beijing, 100080, P. R. China; email: yongrui@microsoft.com.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2014 ACM 1551-6857/2014/02-ART1 \$15.00

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

1. INTRODUCTION

The use of social messaging and broadcasting applications such as microblogs has grown remarkably in the past several years. These microblogs are short messages with a maximum length of 140 characters, and often embed pictures to accompany or supplement the main message. The amount of multimedia-rich messages is poised to grow, especially with better tools that seamlessly integrate social messaging with media uploading and displaying [Miller 2010]. Such a real-time influx of multimedia-rich messages that contain rich semantics brings opportunities to endow images with higher-level semantics beyond object labels. However, this arises new challenges for understanding the association between multimodal multimedia content in multimedia-rich microblogs.

Analyzing the association between pictures and words on microblogs is a new research conundrum. Traditional non-microblog works mainly focus on either image annotation or text illustration alone, and therefore cannot solve all associations between pictures and words in multimedia-rich microblogs. These works assume either that explicit association between visible parts of the picture with nouns and verbs [Wu et al. 2011; Van Zwol and Garcia Pueyo 2012], or that a picture is an illustration of a document with many words [Joshi et al. 2006]. However, *disobeying the fundamental assumptions of traditional annotation, tagging, and retrieval systems, the pictures and words in multimedia-rich microblogs are loosely associated and a correspondence between pictures and words cannot be established*. As a result, the association between pictures and words in microblogs can hardly be established using traditional methods. This underpins the reason for proposing a model with flexible association between images and texts in this paper.

To solve this problem we face the following challenges:

- *Content variety*. In traditional tagging systems, the tags are mostly object labels and the vocabulary size is quite limited. In microblogs, however, the scale of words can reach the entire natural language vocabulary, and even nouns that include both objects and abstract concepts. The content variety makes it very challenging to analyze the association between images and texts in microblogs.
- *Image-text association variety*. In microblogs, the association between images and texts is complex. Some texts express the story depicted by images, and some images illustrate the textual semantics. Moreover, images and texts are not explicitly associated as in an image tagging scenario. While traditional content provider usually has only one of the above association scenarios. Web pages and text books have one or a few images illustrated by a long paragraph of text, while albums have one or a few text titles including many pictures. Both of the conditions only have one-directional correspondence.

To address the above challenges, we propose a novel association model, called Bilateral Correspondence Latent Dirichlet Allocation (BC-LDA) where both image-to-text and text-to-image correspondences are incorporated. In the continuum of different amount of text with respect to pictures in online social multimedia documents, we use a data-driven approach to observe that exploring the mid-point, i.e., loose associations between short social messages and one picture, will enlarge the vocabulary amenable to visual association and give insights about social events. Based on the observation we then design the BC-LDA association model. Our association model considers full correspondence, instead of the previous works using multimodal LDA and correspondence LDA. We demonstrated the power of this model by microblog illustration (i.e., suggesting images to illustrate the text microblogs), image annotation using microblog vocabulary, and multimodal topic exploration over time. On a large data set from real Tencent microblog (one of the largest microblogging services in China [Staff 2011]), we

conduct extensive experiments and demonstrate that our algorithm can significantly and consistently outperform baselines.

The main contributions of this paper can be summarized as follows.

- We propose the first work of understanding multimedia-rich social messages from the perspective of association modeling, which can provide richer content than just using text or image alone.
- We propose a novel Bilateral Correspondence LDA model, which can well address the problem of association modeling in multimedia microblog data to discover both text-to-image and image-to-text correspondence.
- We apply the model to multiple scenarios, including multimodal topic modeling in microblogs, image annotation with microblog vocabulary, and text illustration with microblog images. We extensively evaluate the proposed method in large-scale real microblog data, and demonstrate that the proposed method outperform other baselines in all of these application scenarios.

The rest of this paper is organized as follows: Related works are surveyed in Section 2, and the multimedia-rich data is analyzed to give some observations in Section 3. Section 4 introduces the model, and evaluations are given in Section 5. Conclusions are made in Section 6.

2. RELATED WORK

Our work is part of a long history of research on retrieving images and words, as well as the active research problems in modeling social media.

The relationships between pictures and words can be posed as a bi-directional retrieval problem [Jeon et al. 2003]: *image retrieval or recommendation*, i.e., returning a list of images given one or more words, and *image annotation*, i.e. retrieving one or more words given an image. Formulations and solutions to these problems have since appeared in many interesting ways. Barnard et al. [Barnard et al. 2003] match images and words using statistical topic models, where word similarity is measured by co-occurrence, and image similarity is a segmentation-based approach. The ImageNet [Deng et al. 2009] and NUS-wide [Chua et al. 2009] corpus are among the latest datasets for realistic images from a large-scale web environment, with one or a few tags associated with each picture. Sigurbjornsson et al. [Sigurbjornsson and Van Zwol 2008] use collective knowledge in a social platform to help infer better tags. Shi et al. [Shi et al. 2012] exploits the visual word co-occurrence for image retrieval. These scenarios fall into the “few words per image” category. On the other end of the word-to-image-ratio scale, the story picturing engine [Joshi et al. 2006] retrieves representative images for topical keywords in a large body of text; Cui et al. [Cui et al. 2010] uses content to enrich text features for video classification; the multimedia answering system [Nie et al. 2011] enriches text answers with image and video, retrieved from the web using a collection of answer classification, query generation and image/video re-ranking techniques; Noh et al. [Noh et al. 2009] build a translation engine for multimedia tags based on co-occurrence and online social context. Qi [Qi et al. 2012] corrects imperfect tagging with multi-view learning to achieve better image understanding. These scenarios are illustrative of the “many words per image” category. In recent years, there have been a few works about image tagging and recommendations focusing on social media. Van Zwol [Van Zwol and Garcia Pueyo 2012] proposed a spatially-aware image retrieval method. Wu [Wu et al. 2011] investigated distance metric learning for image tagging. San Pedro [San Pedro et al. 2012] leverages user comments for image re-ranking. Qi [Qi et al. 2011] introduces a semantic knowledge propagation from text corpus to web images. These scenarios are illustrative of the social media category, but they only focus on one direction of the bi-directional retrieval problem.

The statistical models with which to describe images and words has evolved into several types over the past ten years. Language model [Jeon et al. 2003] is a popular choice due to its simplicity and scalability, and is used in systems with many words and few images [Nie et al. 2011; Joshi et al. 2006].

Topic models [Blei et al. 2003], or Latent Dirichlet Allocation (LDA), is the prevailing choice when the task involves modeling the co-occurrences among words and among image descriptors. Observing that an image often occurs multiple times in different microblogs, the file name of an image is treated as equivalent to a text word, and the statistical co-occurrence can be modeled between distinct images and words with LDA, which is the first baseline of this paper. Extensions of topic models has taken into account the distribution of images [Barnard et al. 2003], the temporal topic evolution [Blei and Lafferty 2007] and the explicit correspondence between individual words and image regions [Blei and Jordan 2003]. In all the topic models, word similarity is measured by co-occurrence, and image similarity is based on image features. In our work, we apply SIFT-based approach for its effectiveness in many applications. Li et al. [Li et al. 2010] further develop extended topic models for image regions, categories, and topics. Our work is on the loose correspondence between a few images and words, so we cannot assume that each image descriptor “emits” a word (or vice versa). Therefore the multimodal LDA [Barnard et al. 2003] and correspondence LDA [Blei and Jordan 2003] are the closest approaches, whereas models with stronger assumption about image-word association do not apply [Li et al. 2010]. Multi-modal LDA [Barnard et al. 2003] improves upon the simple co-occurrence LDA by taking into account image similarity. The image is now represented as a bag-of-visual-words, and used in parallel with the text words, which is the second baseline of this paper. The text words and visual words are independently generated in this process. Although the visual similarity of image content is considered and the co-occurrences between text words and visual words are exploited, the correspondences between text words and visual words cannot be inferred from this model because of its assumption of text words and images generated independently. Correspondence LDA [Barnard et al. 2003] improves upon multimodal LDA by including explicit correspondence from image visual words to text words, which is the third baseline of this paper. Before generating a text word, a variable is drawn from uniform distribution to decide which visual word this word should correspond with, and the word distribution depends on both the topic and the correspondence. In this model, the image content plays a more important role than in multimodal LDA, but this model is basically image-driven, i.e., only the correspondences from image visual words to text words are considered, but text words cannot direct the generation of image visual words. There are other works that extend correspondence LDA by modeling the user’s perspective [Chen et al. 2011], by modeling bi-lingual documents [Fukumasu et al. 2012], or by mining topic connections [Chen et al. 2010], but they only focus on image tags or word correspondence and ignore the text to image mapping.

Microblog platforms has been an very active subject of research, from social oriented image search [Liu et al. 2014] to analysis of image tweets [Chen et al. 2013], to social recommendation [Jiang et al. 2012], to using text to predict emotion [Yang et al. 2013], to large-scale heterogeneous hashing [Ou et al. 2013]. A supervised topic model has been used on microblogs [Ramage et al. 2010] to visualize themes in a microblog stream, and observations about topics and users were presented [Qu et al. 2011] for Sina Weibo (another major Chinese microblog service) after a major earthquake in China. Our work is the first to present content analysis on image-enriched microblogs.

3. OBSERVATIONS FOR MULTIMEDIA-RICH MICROBLOG DATA

We collect microblog data from Tencent microblog¹ platform, one of the largest Twitter-style microblog platforms in China [Staff 2011]. To better understand the microblog data, we take note of the following observations.

¹ <http://t.qq.com>

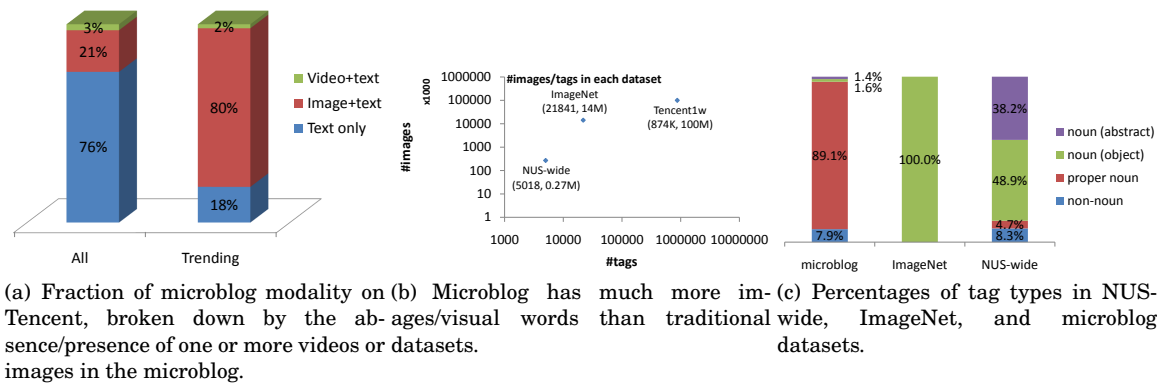


Fig. 1: Differences between microblog and traditional datasets.

Popular microblogs have more multimedia content. We investigate the multimedia ratio on Tencent using hundreds of millions of random data. Within all 719 million tweets posted 6 : 00 Oct 5, 2011 to 12 : 00 Oct 12, 2011 (as part of the experimental dataset, see Section 5), there are about 151 million tweets containing an image, 20.1 million containing a video and the rest are text-only tweets. For more insights on how multimedia microblogs perform, we then collect statistics on the trending (i.e., the most popular) microblogs during 30 minutes from 14 : 00 on Jun 26, 2012, via an API on the Tencent platform. Among the 1000 trending microblogs collected, about 80% include images accompanied with text. Figure 1(a) shows the statistics on all and trending microblogs on the Tencent platform separately. Numbers of text-only microblogs, text-image and text-video microblogs are counted to show their different impacts on the popularity of the microblogs. We observed from the data that popular microblogs have more multimedia content, so that multimedia data analysis is more important for popular microblogs.

Microblogs have many and various tags. In traditional image datasets, thousands of visual words with millions of images are collected. The words are mainly nouns; the images are mainly clear iconic images; and the associations between texts and images are clear. In contrast, the number of words in a microblog can reach the real vocabulary size in natural language, and tens of millions of user-generated images are uploaded during a week. Furthermore, traditional image datasets only have one or several tags for an image, but microblogs usually have more. Figure 1(b) counts the number of visual words and images in NUS-wide, ImageNet, and microblogs during a week. Here image tags are synsets in ImageNet, Flickr tags in NUS-wide, and microblog text words in the microblog dataset. We can observe that microblogs have ten times the size of a traditional dataset with only a week's data. Microblog tends to have more tags for each image. In traditional datasets, tags are always defined within a narrow scope. Nouns which stand for objects accompany several images. In microblog, the tag variety is the same as the vocabulary of natural language. To illustrate this, we check the tags in wordNet [Miller 1995] to classify them into four categories: non-nouns (verbs, adj, etc.), nouns standing for objects, abstract nouns (such as love), and proper nouns (not in wordNet, such as Chinese names). We take the majority of attributes when some tags have multiple attributes. The statistics are shown in Figure 1(c). Traditional datasets have more object nouns, while microblogs have more proper nouns. On one hand, this phenomenon makes modeling the association between images and text more challenging. On the other hand, if we can successfully build a model for microblog data, we can get image tags beyond objects, which cover proper nouns, adjectives, and even verbs.

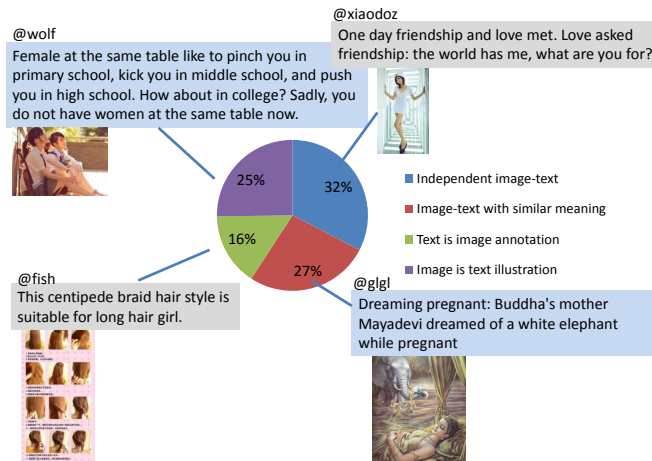


Fig. 2: Examples of the four types of image-text association, and their relative percentage on the Tencent microblog service. See Section 3 for discussions.

Image-text association is important but challenging. In microblogs, image-text association is complex. In Figure 2, we show both statistics and examples for each condition. We randomly collect 200 microblogs with both image and text and carry out a user study to categorize these microblogs into different association conditions. The statistics shows that there are four main association formats: 1) Independent image-text. The author posts an image only for aesthetic purposes. The example microblog talks about love and friendship, but the image is only a beautiful girl which is independent from the text. 2) Image-text with similar meaning but not tight association. The image and text can express the meaning separately, so the association is loose. The example microblog talks about the white elephant and pregnancy in both the image and the text. 3) Text depending on image, i.e., text is image's annotation. The image illustrates the meaning of the content, but the text cannot convey the intended meaning by itself without the combination with support in another medium, which we called "self-illustrated". 4) Image depending on text, i.e., image is text's illustration. The text says everything, but the image cannot be self-illustrated, which is only a supplement of the text. In the experiment section, we have more examples showing that the association is various. In order to model the data well, we need to capture all of these conditions of the association, which is important and challenging for our work. Compared to the well known problem of image annotation [Everingham et al. 2011] or image retrieval, the association between the text and image modality are more flexible in microblog data. This motivates us to design a statistical model that accounts for such bi-directional relationship.

4. THE BILATERAL CORRESPONDENCE LDA MODEL

To demonstrate the value and specialty of multimodal microblog data, it is essential to establish the associations and find conditional relationships between images and texts based on good low-dimensional representations, which is commonly addressed in generative models, such as latent dirichlet allocation (LDA) and its variants.

Table I defines the main notations for a number of such models discussed below. Without loss of generality, we focus on modeling microblogs that have one accompanying image. Hence the number of images and documents are assumed to be the same.

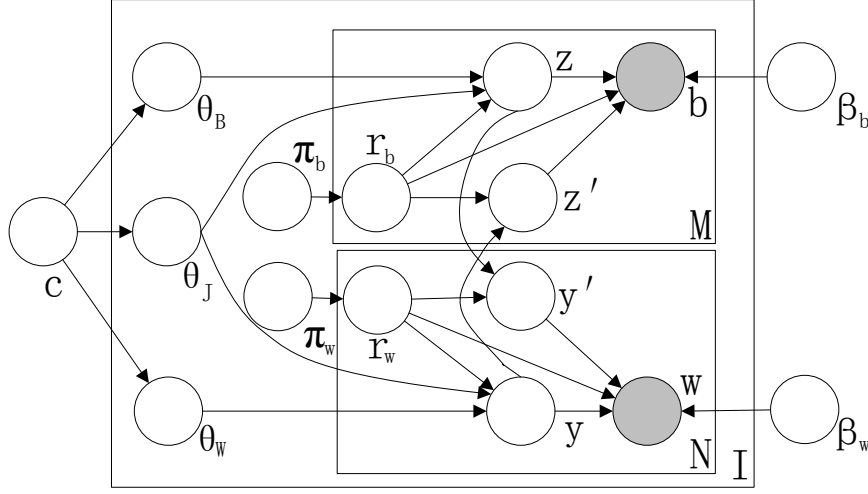


Fig. 3: Bilateral Correspondence LDA. Notations are in Table I, descriptions are in Section 4.

Table I. : Notations for BC-LDA in Microblogs

Notation	Definition
I	the number of microblog documents
N	the number of words in a document
M	the number of visual words in a document
T, T_J, T_B, T_W	the number of topics
$\Phi_t^b, \Phi_t^w, \Psi_t^b, \Psi_t^w,$	model parameters
w	a word
$\mathbf{w} = w_{1:N}$	all words in an image
b	a visual word
$\mathbf{b} = b_{1:M}$	all visual words in an image
z, y	a latent topic of visual words and/or text words
r_m^b, r_n^w	variable for word and visual word correspondence
n, m	indexes to words and visual words
$\alpha_c, \beta^w, \beta^b, \theta_J, \theta_B, \theta_W$	topic model hyper-parameters

4.1 Bilateral Correspondence LDA

We propose a variant of topic modeling to simulate the generation process by integrating various image-text relationships. Correspondence LDA [Barnard et al. 2003] contains a key assumption that all words are generated from an annotation process, i.e., tags are used to exclusively describe the image contents. On microblog data, however, it is common that an image is applied to illustrate the text part, as observed in Section 3. That is to say, for some cases we should first generate the text from the topics and then the corresponding images. This motivates us to propose Bilateral Correspondence LDA in order to allow for bi-directional association of the two modalities. A graphical model of BC-LDA is shown in Figure 3, which we learn with Gibbs sampling [Casella and George 1992] and approximate inference. Compared with correspondence LDA [Barnard et al. 2003], where the relationship only goes from visual word to text word, our BC-LDA has a much more general assumption. Both directions of cross-modal correspondence are included.

We define a set of T multimodal topics, out of which T_J are *joint* topics on visual words and text words with prior θ_J and multinomial distribution Φ_t^b , and Φ_t^w for visual words and text words, respectively. Additional T_B topics are only defined on visual words (or equivalent of emitting a “null” word) with prior θ_B and multinomial visual word distribution Ψ_t^b , and the rest of T_W topics are only defined on words (or equivalent of emitting a “null” visual word) with prior θ_W and multinomial word distribution Ψ_t^w . With $T = T_J + T_B + T_W$. The prior for Φ_t^b and Ψ_t^b are β^b , and the prior for Φ_t^w and Ψ_t^w are β^w .

To allow for flexible correspondence between text words and visual words among these three types of topics, we define a 3-way correspondence variables $r_m^b \in \{0, 1, -1\}$ and $r_n^w \in \{0, 1, -1\}$ for each visual word and each word in the document. Here $r_m^b = 0$ means visual word topic z_m is generated directly from prior and not tied to those of a word; $r_m^b = 1$ means that b_m is tied to (the topic of) one of the existing words, and z_m will be sampled uniformly from $y_{1:N}$ in a manner similar to correspondence LDA; $r_m^b = -1$ means that z_m is one of the visual word-only topics, and b_m is sampled from $\Psi_{z_m}^b$. The reverse holds for r_n^w and w_n . The conditions of r^b and r^w cover each condition in Figure 2. $r^b = 1$ indicates a visual word is generated with a text prior, which is to say that an image is text illustration. $r^w = 1$ indicates a word is generated with a visual word prior, i.e., a word is image annotation. $r^b = -1$ and $r^w = -1$ indicate that the visual words are independent with the words. $r^b = 0$ and $r^w = 0$ indicate that the word and image talk about the same meaning, but without a strict prior relationship.

The generative process is as follows:

- (1) For each of the T_J joint topics,
sample $\Phi_t^b \sim \text{Dirichlet}(\beta^b)$ and $\Phi_t^w \sim \text{Dirichlet}(\beta^w)$.
- (2) For each of the T_B visual word topics and T_W word topics,
sample $\Psi_t^b \sim \text{Dirichlet}(\beta^b)$ and $\Psi_t^w \sim \text{Dirichlet}(\beta^w)$.
- (3) For each of the D documents
 - (a) Sample topic proportion $\theta_d \sim \text{Dirichlet}(\alpha_c)$
 - (b) Sample correspondence priors $\pi_d^b \sim \text{Dirichlet}(\gamma^b)$, $\pi_d^w \sim \text{Dirichlet}(\gamma^w)$.
- (4) For each of the D documents (subscript d omitted for brevity)
 - (a) For each of the M visual words:
Sample a correspondence variable

$$r_m^b \sim \text{Multinomial}(\pi^b).$$

Sample a topic

$$z_m \sim \begin{cases} \text{Multinomial}(\theta_J), & \text{if } r_m^b = 0 \\ \text{Uniform}(y_{(1)}, \dots, y_{(N)}) & \text{if } r_m^b = 1 \\ \text{Multinomial}(\theta_B), & \text{if } r_m^b = -1 \end{cases}$$

Sample a visual word

$$b_m \sim \begin{cases} \text{Multinomial}(\Phi_{z_m}^b), & \text{if } r_m^b = 0 \text{ or } 1 \\ \text{Multinomial}(\Psi_{z_m}^b) & \text{if } r_m^b = -1 \end{cases}$$

- (b) For each of the N words:
Sample a correspondence variable

$$r_n^w \sim \text{Multinomial}(\pi^w)$$

Sample a topic

$$y_n \sim \begin{cases} \text{Multinomial}(\theta_J), & \text{if } r_n^w = 0 \\ \text{Uniform}(z_{(1)}, \dots, z_{(M)}) & \text{if } r_n^w = 1 \\ \text{Multinomial}(\theta_W), & \text{if } r_n^w = -1 \end{cases}$$

Sample a word

$$w_n \sim \begin{cases} \text{Multinomial}(\Phi_{y_n}^w), & \text{if } r_n^w = 0 \text{ or } 1 \\ \text{Multinomial}(\Psi_{y_n}^w) & \text{if } r_n^w = -1 \end{cases}$$

Here z_m and y_n stand for the topic that generate visual word b_m or text word w_n , respectively, which is not an indexed topic. Note that the uniform distribution for the topic correspondence step in 4(a), i.e., sampling z_m when the corresponding $r_m^b = 1$ is not based on the final values of $y_{1:N}$, but based on the current values $y_{(1):(N)}$. In practice, topics are first randomly assigned to every word, and a more appropriate topic is then re-assigned to each word during sampling. Here the latent topics z_m will be based on the current values $y_{(1):(N)}$. The same holds for words in step 4(b).

The joint probability of a document is

$$p(\theta, \mathbf{z}, \mathbf{y}, \mathbf{r}^w, \mathbf{w}, \mathbf{r}^b, \mathbf{b}) = p(\theta|\alpha_c) \prod_{m=1}^M p(r_m^b, z_m, b_m|\theta, \beta^b, \pi^b) \prod_{n=1}^N p(r_n^w, y_n, w_n|\theta, \beta^w, \pi^w) \quad (1)$$

where

$$p(r_m^b, z_m, b_m|\theta, \beta^b, \pi^b) = p(r_m^b = 1|\pi^b)p(b_m|z'_m, \beta^b)p(z'_m|y_{(1):(N)}) \\ + p(r_m^b = 0|\pi^b)p(b_m|z_m, \beta^b)p(z_m|\theta_J) + p(r_m^b = -1|\pi^b)p(b_m|z_m, \beta^b)p(z_m|\theta_B) \quad (2)$$

$$p(r_n^w, y_n, w_n|\theta, \beta^w, \pi^w) = p(r_n^w = 1|\pi^w)p(w_n|y'_n, \beta^w)p(y'_n|z_{(1):(M)}) \\ + p(r_n^w = 0|\pi^w)p(w_n|y_n, \beta^w)p(y_n|\theta_J) + p(r_n^w = -1|\pi^w)p(w_n|y_n, \beta^w)p(y_n|\theta_W) \quad (3)$$

The model first generates the correspondence variables r , which determines whether a visual/text word is generated direct from prior, tied to a text/visual word, or from a single modal topic with only visual/text words. Then visual words and text words are generated with this prior correspondence. Therefore, in this model we consider not only text-visual word correspondence, but also text-visual word independent conditions. This model exploits flexible correspondence relationships between images and texts, which is intuitive and natural in shared online media. This model is similar to a recent cross-lingual topic model [Fukumatsu et al. 2012], with the difference being that our model takes into account topics that exist in one modality only. We will empirically demonstrate the importance of such flexible correspondence relationships between images and texts in Section 4.2.

4.2 Model Inference

Since exact probabilistic inference for Bilateral Correspondence LDA is intractable, we carry out variational inference to approximate the posterior distribution over the latent variables given text/image. In particular, we define the following factorized distribution on the latent variables:

$$q(\theta, \mathbf{z}, \mathbf{y}) = q(\theta|\Omega) \prod_{m=1}^M (q(r_m^b \geq 0|\pi_m^b)q(z_m|\Phi_{z_m}^b) + q(r_m^b = -1|\pi_m^b)q(z_m|\Psi_{z_m}^b)) \\ \prod_{n=1}^N (q(r_n^w \geq 0|\pi_n^w)q(y_n|\Phi_{y_n}^w) + q(r_n^w = -1|\pi_n^w)q(y_n|\Psi_{y_n}^w)) \quad (4)$$

The free variational parameters Ω , Φ , and Ψ can be approximated by minimizing the KL-divergence between this factorized distribution and the true posterior. We give the following algorithm:

(1) Update the posterior Dirichlet parameters:

$$\Omega_i = \alpha_i + \sum_{m=1}^M (q(r_m^b \geq 0 | \pi_{mi}^b) \Phi_{mi}^b + q(r_m^b = -1 | \pi_{mi}^b) \Psi_{mi}^b) + \sum_{n=1}^N (q(r_n^w \geq 0 | \pi_{ni}^w) \Phi_{ni}^w + q(r_n^w = -1 | \pi_{ni}^w) \Psi_{ni}^w) \quad (5)$$

(2) For each topic,

$$\pi_{ni}^w \propto p(y_n = i | r_n^w, \beta^w) \quad (6)$$

$$\pi_{mi}^b \propto p(z_m = i | r_m^b, \beta^b) \quad (7)$$

(3) For each text word,

$$\Phi_{ni}^w \propto p(w_n | y_n = i, r_n^w, \beta^w) \quad (8)$$

$$\Psi_{ni}^w \propto p(w_n | y_n = i, \beta^w) \quad (9)$$

(4) For each image visual word,

$$\Phi_{mi}^b \propto p(b_m | z_m = i, r_m^b, \beta^b) \quad (10)$$

$$\Psi_{mi}^b \propto p(b_m | z_m = i, \beta^b) \quad (11)$$

With the approximate posterior, we can find $p(w|B)$ and $p(b|W)$ as follows:

$$p(w|B) \propto \sum_{y_n} \sum_{m=1}^M (q(r_n^w \geq 0 | \pi_n^w) q(y_n | \Phi_n^w) + q(r_n^w = -1 | \pi_n^w) q(y_n | \Psi_n^w)) p(b | z_m, \beta^b) \quad (12)$$

$$p(b|W) \propto \sum_{z_m} \sum_{n=1}^N (q(r_m^b \geq 0 | \pi_m^b) q(z_m | \Phi_m^b) + q(r_m^b = -1 | \pi_m^b) q(z_m | \Psi_m^b)) p(w | z_n, \beta^w) \quad (13)$$

4.3 Applications of BC-LDA

Based on the estimated parameters and probabilities, we then infer two important conditional probabilities to be used for microblog text illustration and image annotation applications.

Image Annotation. The joint probability of a word w and an image B is calculated by

$$p(w, B) = \sum_z p(z) p(w, B | z) + \sum_y p(y) p(w, B | y) = \sum_z p(z) p(B | z) p(w | B, z) + \sum_y p(y) p(w | y) p(B | w, y) \quad (14)$$

where $p(z)$ and $p(y)$ are given as the topic distribution by the model. $p(B | z)$ is defined as the multiplication of all $p(b | z)$ with $b \in B$. $p(w | y)$ is given by the model with text-only topics. $p(w | B, z)$ and $p(B | w, y)$ are given by the model with multimodal correspondence. Given these probabilities, we can rank words for the given image, and use the top ranked words as the annotations of the image. Media-rich microblog posts that the model learns from and the more sophisticated image-text relationships guarantee the advantages of the proposed method in image annotation application, which are empirically demonstrated in the experiments.

Microblog Illustration. The joint probability of a visual word b and all words W from a microblog post can be calculated by

$$p(b, W) = \sum_z p(z) p(b, W | z) + \sum_y p(y) p(b, W | y) = \sum_z p(z) p(b | z) p(W | b, z) + \sum_y p(y) p(W | y) p(b | W, y) \quad (15)$$

where $p(z)$ and $p(y)$ are given as the topic distribution by the model. $p(W|y)$ is defined as the multiplication of all $p(w|y)$ with $w \in W$. $p(b|z)$ is given by the model with image-only topics. $p(W|b, z)$ and $p(b|W, y)$ are given by the model with multimodal correspondence. Given these probabilities, we can rank images for the given textual microblog, and use the top ranked images to illustrate the microblogs. This application can greatly reduce the user behavior cost of generating multimedia-rich microblogs. In the past, if users wanted to find a proper image to illustrate texts to make the text microblog post more visually attractive, the common method was to search for images on search engines. Then the manually selected images are referred via URL or uploaded onto the microblog platforms. With the proposed model and the invaluable multimedia-rich microblog resources, we can recommend images for the input text to illustrate it with only one step.

5. EXPERIMENTS

In this section, we discuss the performance of the BC-LDA model with quantitative evaluations on held-out log-likelihood of topics with rich media retrieval. We also present a number of illustrative examples on topic visualization, picture annotation and text illustration.

5.1 Data and Setup

Information Extraction from Text. Each microblog contains one or more Chinese sentences or phrases. We use an off-the-shelf tool to segment a sentence into words [Sproat and Emerson 2003]. This tool is used in the production system in Tencent and performs robustly on word segmentation and word-sense tagging. As stated in Section 3, Microblogs contain words beyond objects, such as names of people, abstract words like *love* or *hero*, and words that have diverse visual appearances, such as *TV show* and *astrology*. We keep the nouns and proper nouns, we also filter out stop words using generic and platform-specific list (e.g. Tencent, since it is the name of the platform thus appears too frequently). The remaining words are very different from tags in traditional image annotation. We have both object nouns and abstract nouns.

Microblog Image Processing. Each unique microblog image is processed in three steps:

- Extracting the scale-invariance feature point (SIFT) and its 128-dimensional multi-scale gradient descriptor [Lowe 1999] for each image.
- Constructing a visual codebook by K-means on all SIFT descriptors from a large sample of diverse images. We then store the cluster centers as the codebook. These centers remain fixed throughout further processing and modeling.
- Each SIFT descriptor (typically numbering hundreds) from each image is then assigned to its closest codebook vector. This yields a V_b -dimension bag of visual words vector [Sivic and Zisserman 2003], which is then normalized, summing to one.

To remove image near-duplicates, we use the nearest-neighbor query on their V_b -dimensional feature vector, and cut it off with a tight threshold. This procedure is tuned to have high level of precision. In this work, we extract bag-of-visual-words from each image, with the visual codebook size $V_b = 500$. We are aware that related research has shown that better results [Moosmann et al. 2007] can be achieved with a larger V_b . But the focus of our paper is to compare different image-word association models under the same visual processing setup, and 500 visual words is enough for this purpose.

Data Collection and Filtering. Our evaluation data set consists of all Tencent microblogs associated with at least one image between 6 : 00 Oct 5th, 2011 and 12 : 00 Oct 12th, 2011, and between 10 : 00 June 20th, 2012 and 15 : 00 June 26th, 2012. There are 395,531,919 such image-microblogs and 308,724,962 images with distinct URLs. The frequency statistics of the image URLs are shown in Figure 4. Very frequent images such as the logo of a game called *Cross Fire* appear many times, and

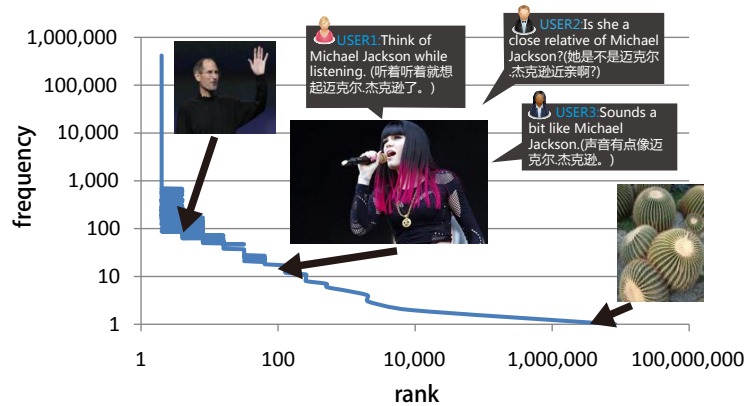


Fig. 4: Image URL frequency vs rank.

plays a role of *stop words*, hence they should be removed. A large number of pictures with low frequency that are of personal nature, such as the cactus photo, are seen only once. Images that appeared a few times, i.e., mid-frequency range are more informative. After visual de-duplication of the images and vocabulary pruning, we removed the most and least frequent images (replies less than 10), keeping 257,301 unique images and 7,555 nouns from the image microblogs. In the down sampling step, we keep microblogs that have different text descriptions within the same image. An example is shown in Figure 4, which reflect different information of the same image. Therefore, if a retweet has exactly the same text as the original tweet, it will be removed in the down sampling; if they provide new information, they will be kept for the image. Such duplicate removal in microblog text ensures that multiple retweets do not skew or dominate word statistics.

We implement the system on one PC with a 2.50GHz CPU and 3.25GB of memory. Training the model to learn parameters with 762,438 documents, 150 topics, and 100 iterations takes about 20 hours. That is to say, we can process one week of data within one day, which is acceptable for most large-scale online applications.

5.2 Multi-modal Topics Evaluation

To evaluate our multi-modal topics, we calculate objective held-out log-likelihood, but also show real examples of these topics.

Held-out log-likelihood. We can evaluate the quality of each topic model by measuring the held-out log-likelihood. A higher held-out log-likelihood indicates better predictive abilities of the model. Thus we estimate the topic models with the training set and compute the log-likelihood of generating the held-out set. Table II shows the held-out log-likelihood of each topic model estimated with the microblog data with the number of topics varying from 30 to 210. CorrLDA has the best performance when the total topic number is 30. BC-LDA performs best for all larger topic numbers. This is because when the topic number is small, the three conditions for BC-LDA have topic numbers that are too limited for each case, and thus limit the model's generating ability. The overall result demonstrates that BC-LDA performs at a superior level compared to the baselines.

Multi-modal Topic Examples. We examine a few representative joint topics with their words and top images in Table III. We include a short summary of the topic theme, the topic type, the top words (in Chinese with their English translation) and the top three distinct images from the top-scoring documents related to each topic. We can see that these topics are meaningful in both words and pictures.

Table II. : Per-word held-out log-likelihood. Boldface indicates the best result in each column.

Model	Topics						
	30	60	90	120	150	180	210
Multimodal LDA	-26.624	-19.457	-16.065	-14.184	-13.033	-12.215	-11.602
CorrLDA	-18.0512	-18.835	-17.2906	-14.170	-12.987	-12.166	-11.516
BC-LDA	-22.1667	-18.219	-15.509	-13.688	-12.527	-11.694	-11.106

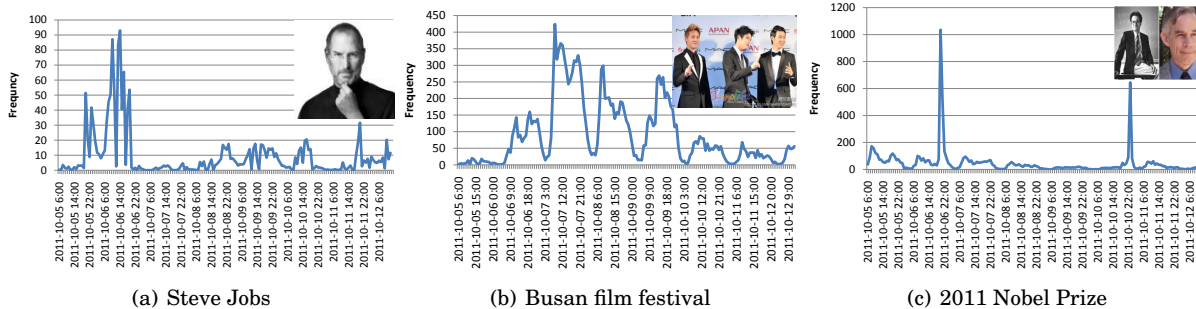


Fig. 5: The volume of three sample topics over time.

The topic categories are quite broad, including objects (e.g., food), events (e.g., the passing of Steve Jobs and group-buying) and some abstract concepts (e.g., hero). Words and images in the topics *group buying* (Topic 1 in Table III) and *Steve Jobs* (Topic 2 in Table III) can separately illustrate the content of topic, thus the topics generate images and texts without prior requirements. The topics *game* (Topic 3 in Table III) and *Euro Cup* (Topic 4 in Table III) are mainly based on the texts, while the images cannot make the meaning clear by themselves, thus these topics generate the images with a prior text. The topic *cellphone* (Topic 5 in Table III) is mainly based on images, while the texts depend on the prior image. The topics *Dragon boat festival* (Topic 6 in Table III) and *news of Diaoyu Islands* (Topic 7 in Table III) are single modal topics with only words, because the accompanying images are usually unrelated to the topics. The topic *landscape* (Topic 8 in Table III) is a single modal topic with only images, the text being independent. To illustrate the prior image or prior word, we carry out pure image clustering and pure text clustering. The results are shown in Table IV. Without the prior text such as *Euro Cup* or *Italy*, green leaves are similar to the soccer ground for their similar color. Without the prior image for cellphones, words such as *computer* and *camera* will be considered similar with other real topic words. Therefore, the prior image and prior text in a joint topic plays an essential role in generating these topics.


Figure 5 plots topic volume (sum of posterior over all documents) for each hour over the seven days after 6 : 00 Oct 5th, 2011. Topics related to three international events are manually selected for this illustration. The Steve Jobs topic (Figure 5(a)) peaks on October 6th, the day of his passing. When the Busan Film Festival (Figure 5(b)) started on October 6th and ended on October 14th, we can see several daily post cycles from the trace. There were two Nobel prizes announced during that week. The first topic peak on October 7th Beijing time corresponds to the Nobel Prize in Literature and the second one on October 11th is in Economics (Figure 5(c)).

Online topic discovery system. The most practical application of our algorithm is to discover multimedia topics from multimedia-rich microblog. Besides general topic discovery on full data, our BC-LDA model can also handle sub-topic discovery in a specific domain [Wang et al. 2012]. With a few words given by the search entry, we can filter the data to a narrowed domain. By applying our model on

Table III. : Topic Examples from image-rich microblogs. One or two examples are collected to show here for each case of the topics and the associations.

Topic Type	Top Words	Top Images	Topic ID & Summary
Joint topic with separate image and text	group-buying (团购) original price (原价) miracle (奇迹) nationwide (全国)		1. Advertisement for group-buying
	Jobs (乔布斯) Apple (苹果) corporation (公司) Steve Jobs (史蒂夫·乔布斯)		2. Celebrity
Joint topic with image as text illustration	heros (英雄) honor (荣誉) warloads (群雄)		3. Game
	Europe (欧洲) Germany (德国) Italy (意大利) soccer (足球)		4. Europe Cup
Joint topic with text as image annotation	technology (科技) digital (数码) cellphone (手机) Apple (苹果)		5. Cellphone
Single modal topic with only words	Zongzi (rice dumplings, 粽子) dragon boat festival (端午节) dragon boat (龙舟)		6. Dragon boat festival
	Diaoyu Islands (钓鱼岛) warship (军舰) ministry of foreign affairs (外交部) territory (领土)		7. News of Diaoyu Islands
Single modal topic with only images			8. Landscape

Table IV. : Image/text clustering without priors. This is carried out separately from our model as baselines to show case when priors do not exist.

Pure-image clustering	Pure-text clustering
	Technology (科技), Internet (互联网), digital (数码), cellphone (手机), computer (计算机), Apple (苹果), camera (相机), iPad,

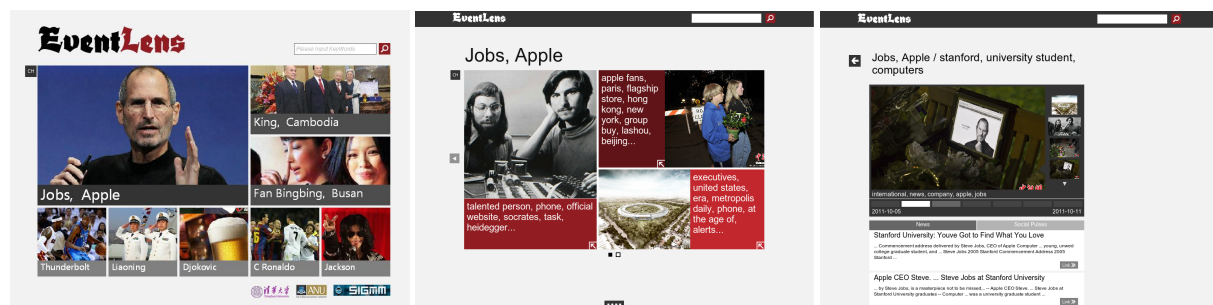


Fig. 6: Online magazine system.

this filtered data, we discover sub-topics to show details. We have built a system² and a demo video³. The demo video shows some cases to illustrate the performance. With the model, each event has its corresponding texts and images. In this paper, we showcase three weeks' microblog data to build the system and retrieve events. This system not only relies on high image annotation accuracy, but also requires accurate text-image relevance. Figure 6 contains several screenshots of the digital magazine system, including several trending topics with images and their top keywords. The left screen shows an overview of one week's trending events, each with a top image and several top topic words. The middle screen shows a summary of several sub-events in one specified event, each with one or two images with more topic words. The right most screen shows details of one specified sub-event, with top images, top user microblogs, and related news items. This demo demonstrates that our model not only performs well on the static dataset, but also provides a great user experience with online data by showing multiple facets of the events, including topic words, users, microblogs, web news, and images with social tags.

5.3 Microblog illustration with images

One application for the BC-LDA model is to recommend pictures to illustrate a text-only microblog. Specifically, we use 80% of the dataset for training a topic model, 10% held-out data for topic evaluation and the other 10% is used for testing. The words in the test microblog are used as input, and a ranked list of images serve as the output. We evaluate this list using two different sets of ground truth. The *strict* ground truth regards the picture that actually appeared with the input microblog (and its near-duplicates) as the correct entry and the rest as incorrect. The *subjective* ground truth is obtained by

²<http://socialmedia1.3322.org/>

³<https://www.dropbox.com/s/xqf3wmdm68l66ev/doubleblinddemo.wmv>

taking 10 microblogs from each of 10 random topics⁴. The union of the top 5 returned pictures are pooled from all model variants, and three human evaluators judge whether or not a picture is semantically relevant to the microblog. We construct the *subjective* ground truth since typically more than one pictures can be used to illustrate a particular microblog, and the *strict* ground truth cannot capture such diversity.

Given a microblog W (parsed into words), we estimate the probability of a picture B using Equation 15. All pictures are ranked according to $P(B|W)$, and we use both precision at top K (P@K) and Kendall's Tau to evaluate $P(B|W)$ with respect to the ground truth. The Kendall's Tau distance $K(\tau_0, \tau)$ [Fagin et al. 2003] computes the fraction of pairs that disagree between a returned list and the ground truth. Its value lies in $[0, 1]$, a smaller K means a better retrieval result. In our experiment, a ground truth list τ_0 with n_1 positive entries and n_0 negative entries can be written as an n_1 -by-1 vector of 1s followed by a n_0 -by-1 vector of 0s, i.e., $\tau_0 = [1_{n_1}; 0_{n_0}]$. A returned list of length T can be written as a binary vector

$$\tau = [\tau(1), \tau(2), \dots, \tau(T)] \in \{0, 1\}^T.$$

The Kendall tau distance is then computed as the fraction of 0-1 pairs to be swapped for sorting τ into τ_0 , where both lists are padded to have the same length and the same number of 1s.

We learn 150 topics for BC-LDA and three models. *Strict* and *subjective* evaluations are carried out with 200,000 microblogs, where 1/10 of these microblogs (20,000) are served as test queries, and the rest 9/10 of the microblogs are served as training data. A 10-fold evaluation is done to show the average value and the error bar. Figure 7(b) compares their retrieval performance on the *strict* ground truth over all 20,000 test queries. We can see that the Kendall tau distance from BC-LDA is notably lower than all three baseline models, and it also has the best average Precision@K. Correlation LDA slightly under-performs BC-LDA in P@K, but the co-occurrence model and simple multi-modal model are significantly worse. The *strict* ground truth is only counting visually duplicate images without rewarding many semantically relevant images, leading to the low-precision in Figure 7(b). We also evaluate P@K on the *subjective* ground truth for varying K, as shown in Figure 7(c). The three stronger models BC-LDA, CorrLDA and multimodal LDA are included in this evaluation, we can see that BC-LDA still out-performs the two baseline models by a margin, and its P@5 is over 60%. We have also tuned the number of joint topics to observe its effect on performance. We have trained the BC-LDA model on a subset of 10,000 microblogs, and recorded the Kendall tau distance over 1000 test queries. This experiment on the subset does not sufficiently show the performance on the whole dataset, but it illustrates that our algorithm is not sensitive to the topic number, so that we select 150 as the topic number of our experiment is reasonable. Figure 7(a) shows that best performance is reached for 180 topics, and 150 or 210 are close to optimal. Compared to Figure 7(b), the Kendall Tau values are worse here. That's because they use different data.

Four example microblog illustrations are presented in Table V. The English translation and original Chinese text are included for each query microblog and the top-ranked unique pictures are presented on the right. If the original image appears in the top three, we will mark it with the word "original" on the bottom right corner of the image. Note that the top images are different from that of any single topic, as their ranking is obtained by aggregating all words in the query, all visual words in the result image, and all topics (Equation 15), while Table III contains the top images for one topic only. The confidence of these resulted images are quite high (over 0.9) in our showcase, so that they have a good performance.

⁴The topics include (manually summarized title) *Steve Jobs*, *Group Buy*, *Astrology*, *Car*, *Gaddafi*, *Nobel Prize*, *TV drama*, *Game*, *Na Xie (a hostess)'s wedding*, *Mural (a film)*.

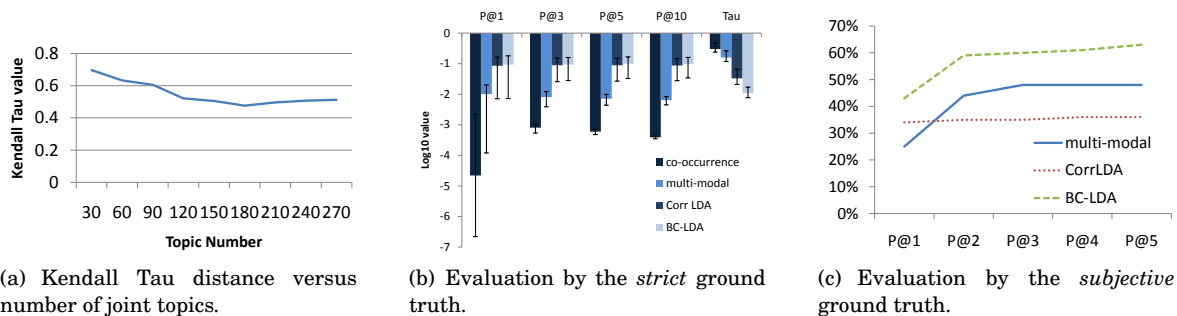


Fig. 7: Microblog illustration performances. Given a text query, an image is predicted by BC-LDA as an illustration of the text (details in section 5.3). Examples of strict and subjective ground truth are in Table V.

Table V. : Microblog illustration examples for some text queries (The original image is marked).

Query	Top Images
The Apple fanse of Beijing Salitun are dedicating flowers to Jobs. (北京三里屯果粉向乔布斯献花悼念-美女过来看)	Original
#StartlingbyEachStep (TV drama) is now on, Maertai Ruoxi is coming again via time-travel. (#步步惊心#火热上映, 马尔泰·若曦再度穿越而来)	Original
A man's beauty is in his depth and sincerity; A woman's beauty is in her charm and expressions. (男人美, 美在深度和真诚; 女人美, 美在风度和表情)	Original
Astrology signs with a rough love life: Cancer, Virgo, ... Astrology signs prone to loving the wrong person: Aries, ... (情路走得很坎坷的星女: 巨蟹座、处女座、天蝎座; 爱情中最容易爱错认的星座: 白羊座、双鱼座、 双子座、摩羯座、水瓶座)	Original

From these evaluations, we can see that BC-LDA consistently out perform the baselines, and the recommendations are sensible. One application scenario can be to present the recommended picture to a user who is about to post on a microblog but does not have proper images. He or she can then choose to post some of the recommended pictures alongside the text part of the microblog entry.

5.4 Image Tagging

Another application of BC-LDA is to explain a given image with words seen in microblogs. This evaluation is done with 10,189 test images. We use words in the original microblog as the ground truth. If a unique image appears in more than one microblog, we count the word frequency among all the microblogs in which the image appears. The 10 most frequent words (stop words excluded) are taken as the ground truth. We estimate the probability $P(w|B)$ of each word w given test image B . We evaluate Precision@K with K varying from 1 to 10, as show Figure 8. Besides evaluating all tags in Figure 8(a), we also pay special attention to abstract nouns in Figure 8(b). The top 1 to 10 abstract nouns are

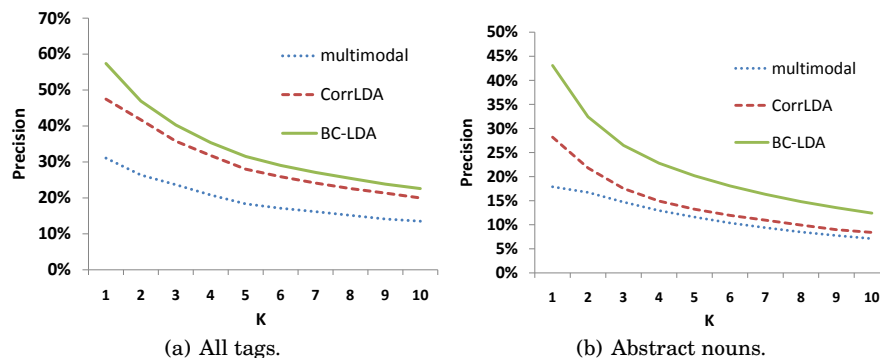


Fig. 8: Image tagging evaluation: precision at top K.

compared with the ground truth. All tags are a bit better than abstract nouns, because object nouns are easier to match with images while abstract nouns are more difficult to establish correspondence with images. In both settings, BC-LDA has the best performance. With all tags, BC-LDA leads correlation LDA with a small margin and multimodal LDA with a large margin, while with abstract nouns, BC-LDA is far better than the other baselines. This is because text-to-image correspondence helps abstract nouns significantly while image-to-text correspondence does not help much. Example assigned tags are shown in Table VI. Note that the overall accuracy for the top-ranked tag is $\sim 60\%$, which is quite high in our scenario. Previous image tagging work usually has 100 to 1000 tag categories, but our tag number can reach the real vocabulary. For the three example images, BC-LDA correctly identifies the person name. It also reveals the relevant event context such as location, the role of the person (e.g. *actress*) and abstract descriptors (*classic*). These images are also found on Flickr, the popular photo sharing website. The Flickr tags in the last column of Table VI are found by querying Flickr using topic keywords in a given date range, and manually identifying visual duplicates with popular microblog photos. We can see that although in some cases Flickr tags are equally informative (*Busan Film Festival*) they can be less context-rich, related to the photo owner and not to the event (*BIFF*) or related to the person but not related to the event (*iTunes*). One reason why BC-LDA performs well is that image tags, assigned by the user or machines, are mainly about image content and production metadata, and less about relevant events.

6. CONCLUSION AND FUTURE WORK

We have proposed a novel bilateral correspondence model for multimedia-rich microblog data. By analyzing the association between multimodal multimedia content in microblog streams, we have discovered multimodal topics from microblogs by establishing correspondences between images and texts in microblogs. The bilateral correspondence model establishes flexible correspondence between pictures and words, including both uni-modal topics and joint topics. Text-to-image and image-to-text correspondence are both considered in the model and play essential role in modeling various types of microblog data. The experiment results show superior performance in microblog illustration and image annotation against the baselines. For further research, microblog analysis could be extended to the individual processing of verbs, adjectives, and proper nouns with different priors, which can better fit the data instead of a global prior for all words in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China, No. 61370022, No. 61003097, No. 60933013, and No. 61210008; International Science and Technology Cooperation Program of China,

Table VI. : Tag comparison between Flickr and BC-LDA. Positive tags are in bold font. Negative tags are gray in color. Neutral tags have normal fonts. Photo credits: Flickr user *COG LOG LAB* shared under CC BY-NC-SA 2.0 <http://www.flickr.com/photos/cogloglab/6216048568/> (top) and Flickr user *crizeel ong* shared under CC BY-2.0 <https://www.flickr.com/photos/crizeelong/8949358694> (bottom).

Image	BC-LDA tags	Flickr tags
	Jobs/乔布斯 , Quotations/格言 , Apple/苹果 , Classic/经典 , University/大学, Hong Kong/香港, God/上帝	Steve Jobs/乔布斯 , Apple/苹果 , iTunes/iTune, App Store/应用商店, iTunes Universal Payment System/iTune 国际支付系统
	Fan Bingbing/范冰冰 , Busan/釜山 , Film Festival/电影节 , Star/明星 , Actor/演员, International/国际, Actress/女星	16th Busan International Film Festival/16 届釜山电影节 , 16th Busan International Film Festival photos/16 届釜山电影节照片, BIFF pictures/釜山电影节照片, BIFF hosted/釜山举办, UhmJi Won pictures in BIFF/ UhmJi 获奖, Ye Ji Won photos in BIFF/Ye Ji 获奖

No. 2013DFG12870; National Program on Key Basic Research Project, No. 2011CB302206; national “1000 People Plan” starting grant. Thanks for the support of NExT Research Center funded by MDA, Singapore, under the research grant, WBS:R-252-300-001-490.

REFERENCES

- K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. 2003. Matching words and pictures. *The Journal of Machine Learning Research* 3 (2003), 1107–1135.
- D.M. Blei and M.I. Jordan. 2003. Modeling annotated data (*SIGIR '03*). ACM, 127–134.
- D.M. Blei and J.D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics* (2007), 17–35.
- D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3 (2003), 993–1022.
- George Casella and Edward I George. 1992. Explaining the Gibbs sampler. *The American Statistician* 46, 3 (1992), 167–174.
- Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. Understanding and classifying image tweets. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 781–784.
- X. Chen, X. Hu, Y. An, Z. Xiong, T. He, and EK Park. 2011. Perspective hierarchical dirichlet process for user-tagged image modeling. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 1341–1346.
- X. Chen, X. Hu, Z. Zhou, C. Lu, G. Rosen, T. He, and EK Park. 2010. A probabilistic topic-connection model for automatic image annotation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 899–908.
- T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. 2009. Nus-wide: A real-world web image database from national university of singapore. In *Proc. Intl. Conf. on Image and Video Retrieval*. ACM.
- Bin Cui, Ce Zhang, and Gao Cong. 2010. Content-enriched classifier for web video classification (*SIGIR '10*). 619–626.
- J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*. Ieee, 248–255.
- M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2011. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. <http://www.pascal-network.org/challenges/VOC>. (2011).
- R. Fagin, R. Kumar, and D. Sivakumar. 2003. Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 28–36.
- Kosuke Fukumasu, Koji Eguchi, and Eric Xing. 2012. Symmetric Correspondence Topic Models for Multilingual Text Analysis. In *Advances in Neural Information Processing Systems* 25. 1295–1303.

- J. Jeon, V. Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models (*SIGIR '03*). 119–126.
- Meng Jiang, Peng Cui, Rui Liu, Qiang Yang, Fei Wang, Wenwu Zhu, and Shiqiang Yang. 2012. Social contextual recommendation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 45–54.
- Dhiraj Joshi, James Z. Wang, and Jia Li. 2006. The Story Picturing Engine—a system for automatic text illustration. *ACM Trans. Multimedia Comput. Commun. Appl.* 2 (February 2006), 68–89. Issue 1. <http://doi.acm.org/10.1145/1126004.1126008>
- L.J. Li, C. Wang, Y. Lim, D.M. Blei, and L. Fei-Fei. 2010. Building and using a semantivisual image hierarchy (*CVPR'10*). IEEE, 3336–3343.
- Shaowei Liu, Peng Cui, Huanbo Luan, Wenwu Zhu, Shiqiang Yang, and Qi Tian. 2014. Social-oriented visual image search. *Computer Vision and Image Understanding* 118 (2014), 30–39.
- D.G. Lowe. 1999. Object recognition from local scale-invariant features (*ICCV*). IEEE, 1150–7.
- George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
- Ross Miller. 2010. Twitter unveils new website with picture and video content embedded on site. (2010). <http://www.engadget.com/2010/09/14/twitter-relaunches-main-site-with-content-embedded-on-site> .
- F. Moosmann, B. Triggs, F. Jurie, and others. 2007. Fast discriminative visual codebooks using randomized clustering forests. (2007).
- Liqiang Nie, Meng Wang, Zhengjun Zha, Guangda Li, and Tat-Seng Chua. 2011. Multimedia answering: enriching text QA with media information (*SIGIR '11*). 695–704.
- Tae-Gil Noh, Seong-Bae Park, Hee-Geun Yoon, Sang-Jo Lee, and Se-Young Park. 2009. An automatic translation of tags for multimedia contents using folksonomy networks (*SIGIR '09*). 492–499.
- Mingdong Ou, Peng Cui, Fei Wang, Jun Wang, Wenwu Zhu, and Shiqiang Yang. 2013. Comparing apples to oranges: a scalable solution with heterogeneous hashing. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 230–238.
- G.J. Qi, C. Aggarwal, and T. Huang. 2011. Towards semantic knowledge propagation from text corpus to web images. In *WWW*. ACM, 297–306.
- Zhongang Qi, Ming Yang, Zhongfei Mark Zhang, and Zhengyou Zhang. 2012. Multi-view learning from imperfect tagging. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 479–488.
- Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. 2011. Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake (*CSCW*). ACM, 25–34.
- D. Ramage, S. Dumais, and D. Liebling. 2010. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*. The AAAI Press.
- J. San Pedro, T. Yeh, and N. Oliver. 2012. Leveraging user comments for aesthetic aware image search reranking. In *WWW*. ACM, 439–448.
- Miaojing Shi, Xinghai Sun, Dacheng Tao, and Chao Xu. 2012. Exploiting visual word co-occurrence for image retrieval. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 69–78.
- B. Sigurbjörnsson and R. Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proc. WWW*. ACM, 327–336.
- J. Sivic and A. Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 1470–1477.
- R. Sproat and T. Emerson. 2003. The first international Chinese word segmentation bakeoff. In *SIGHAN workshop on Chinese language processing*. ACL, 133–143.
- China Internet Watch Team Staff. 2011. Total Weibo Users: Sina v.s. Tencent. (Nov. 2011). <http://www.chinainternetwatch.com/1296/total-weibo-users-sina-tencent> .
- R. Van Zwol and L. Garcia Pueyo. 2012. Spatially-aware indexing for image object retrieval. In *WSDM*. ACM, 3–12.
- Zhiyu Wang, Peng Cui, Lexing Xie, Hao Chen, Wenwu Zhu, and Shiqiang Yang. 2012. Analyzing social media via event facets. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 1359–1360.
- P. Wu, S.C.H. Hoi, P. Zhao, and Y. He. 2011. Mining social images with distance metric learning for automated image tagging. In *WSDM*. ACM, 197–206.
- Yun Yang, Peng Cui, Wenwu Zhu, and Shiqiang Yang. 2013. User interest and social influence based emotion prediction for individuals. In *ACM international conference on Multimedia*. ACM, 785–788.