Electronic Preprint for Journal of Information Processing Vol.24 No.2

Recommended Paper

Bayesian Non-parametric Inference of Multimodal Topic Hierarchies

Takuji Shimamawari[1,a)] Koji Eguchi[1,b)] Atsuhiro Takasu[2,c)]

Received: June 3, 2015, Accepted: December 7, 2015

Abstract: Research on multimodal data analysis such as annotated image analysis is becoming more important than ever due to the increase in the amount of data. One of the approaches to this problem is multimodal topic models as an extension of Latent Dirichlet allocation (LDA). Symmetric correspondence topic models (SymCorrLDA) are state-of-the-art multimodal topic models that can appropriately model multimodal data considering inter-modal dependencies. Incidentally, hierarchically structured categories can help users find relevant data from a large amount of data collection. Hierarchical topic models such as Hierarchical latent Dirichlet allocation (hLDA) can discover a tree-structured hierarchy of latent topics from a given *unimodal* data collection; however, no hierarchical topic models can appropriately handle multimodal data considering inter-modal mutual dependencies. In this paper, we propose h-SymCorrLDA to discover latent topic hierarchies from multimodal data by combining the ideas of the two previously mentioned models: multimodal topic models and hierarchical topic models. We demonstrate the effectiveness of our model compared with several baseline models through experiments with three datasets of annotated images.

Keywords: multimodal topic models, hierarchical topic models, hierarchical clustering

1. Introduction

The use of multimodal data such as annotated images has increased explosively with the growth of social media services, such as Flickr, a photo-sharing community site. Therefore, research on search and analysis of such multimodal data is becoming more important than ever. Topic modeling is one such approach that was originally developed in the field of text analysis [1]. It assumes that each document is represented as a mixture of topics, where each topic is represented as a multinomial distribution over words. Topic modeling was extended to multimodal data such as annotated images [2]. Symmetric correspondence topic models (SymCorrLDA) [3] are state-of-the-art multimodal topic models that can model mutual dependencies between images and the corresponding annotations.

Incidentally, hierarchically structured categories can help users find relevant data from a large amount of data collection. Given a category tree, we can easily find information by tracing a path from the root towards a leaf of the tree. Hierarchical topic models such as Blei et al.'s Hierarchical latent Dirichlet allocation (hLDA) [4] can discover a tree-structured hierarchy of latent topics from a given *unimodal* data collection. Such a tree hierarchy can be used to develop hierarchically structured categories or improve existing category structures. Moreover, hierarchical topic models such as hLDA can generate more accurate topics: some are more general and some others are more specfic, while non-hierarchical topic models such as Latent Dirichlet allocation (LDA) [1] cannot distinguish such generality and specificity of topics. Such accurate topics are expected to bring more powerful prediction power.

Now let us suppose two dog images: one is annotated with "dog" and the other with "puppy." hLDA discovers topic hierarchies by either image features or text annotations. Otherwise, even if the image features and the text annotations are mixed together, the frequency of image features is usually larger than the frequency of text features in such annotated images, and thus the topic hierarchies are mostly based on image features. More recently, Li et al. [5] developed a multimodal hierarchical topic model; however, it uses a strong assumption that only a *one-way* dependency from images to the corresponding annotations is considered. For the previous example, if the image features do not look similar, the dog image is placed apart from the puppy image in the topic hierarchy.

To address the problems above, we propose a hierarchical symmetric correspondence topic model, which we call h-SymCorrLDA, to discover latent topic hierarchies from multimodal data considering the mutual dependencies between images and the corresponding annotations. This model combines the advantages of the two previously mentioned models: multimodal topic models and hierarchical topic models. For the example of the dog images, our h-SymCorrLDA can assign the same topic to annotations of "dog" and "puppy," so the dog image may be placed closer to the puppy image in the topic hierarchy even if

[1] Kobe University, Nada, Kobe 657–8501, Japan
[2] National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] shimamawari@cs25.scitec.kobe-u.ac.jp
[b)] eguchi@port.kobe-u.ac.jp
[c)] takasu@nii.ac.jp

A preliminary version of this paper was presented at the IPSJ Kansai-Branch Convention 2014 held in September 2014. This paper was recommended to be submitted to the Journal of Information Processing (JIP) by the chairman of IPSJ Kansai Branch.

Table 1 Notation used in this paper.

$\mathcal{T}$	Infinite tree
L	Max. level of the tree
D	No. of documents
N_d	No. of word tokens in document d
W	No. of word types
f_i	No. of customers (or documents) at table i
$n_{t,v}$	No. of times when word type v is assigned to topic t
$n_{d,\ell}$	No. of word tokens assigned to level ℓ in document d
$m_{d,k}$	No. of word tokens whose pivot flags specifying mode k in document d

the image features do not look similar. We evaluate our model and some previous models through experiments with three annotated image datasets in terms of test set log-likelihood and normalized mutual information to demonstrate the effectiveness of our model.

2. Related Work

Some researchers have explored topic models for image data [2], [6], [7]. Correspondence topic models (CorrLDA) [2] particularly provide a good theory for modeling the (one-way) dependencies between an image and the text features. More recently, Symmetric correspondence topic models (SymCorrLDA) successfully capture inter-modal dependencies of latent topics in multimodal data. The objectives of those models are to discover *non-hierarchical* topics underlying multimodal data.

In another line of studies, Hierarchical latent Dirichlet allocation (hLDA) [4] was proposed to discover hierarchies of latent topics from *unimodal* data. Li et al. [5] then developed a multimodal hierarchical topic model, assuming a *one-way* dependency from images to the corresponding annotations, similarly to CorrLDA that was previously mentioned.

This is the first work to appropriately discover hierarchical topics underlying multimodal data, considering inter-modal mutual dependencies in the data.

3. Preliminaries

In this section, we first define the notations used in this paper. We also briefly describe the nested Chinese restaurant process (nCRP), which is often used for hierarchical topic models and is fundamental to this study. We then overview Hierarchical latent Dirichlet allocation (hLDA) that is based on nCRP.

3.1 Notation

Following the terminology of topic modeling, we refer to each unit of multimodal data as a *document* that consists of multiple modes. For instance, each annotated image consists of two modes: image mode and text-annotation mode. Moreover, each mode consists of a bag of features, which we refer to as *words*; for instance, there are visual words [6] for the image mode and text words for the text-annotation mode. **Table 1** lists the notations we use in this paper. Note that when an index is replaced with '·', it represents the summation over all possible choices of the index. The subscript '−' indicates when the following element is removed.

3.2 Nested Chinese Restaurant Process

We first review the Chinese restaurant process (CRP) [8], which is a stochastic process that generates a probability distribution on partitions of integers. The CRP with parameter γ, a positive real number, can be described by the following metaphor. Suppose that we have a Chinese restaurant with a countably infinite number of tables. Customers $\{1, 2, \ldots, D\}$ come into the restaurant one at a time and have to choose a table at which to sit. The probability that the d-th customer sits at the i-th table is given by the following distribution:

$$p(c_d = i \mid c_{1:(d-1)}) = \begin{cases} \dfrac{f_i}{d+\gamma-1} & \text{(if } i \text{ is existing)} \\ \dfrac{\gamma}{d+\gamma-1} & \text{(if } i \text{ is new)} \end{cases} \tag{1}$$

where c_d indicates the table at which the d-th customer sits and γ controls the probability that the customer chooses a new table. When D customers sit at tables in accordance with the distribution above, the table assignments of all customers represent a partition of D integers.

The nested Chinese restaurant process (nCRP) [4] generates a probability distribution on trees as an extension of the CRP. The nCRP can be described by the following metaphor. Suppose that there is a countably infinite number of restaurants, each of which has a countably infinite number of tables. One restaurant is identified as the root restaurant, and on each table in the root restaurant, there is a card with the name of another restaurant. A customer first comes into the root restaurant and chooses a table in accordance with the CRP. He or she then moves to the restaurant indicated by the card on that table and chooses a table in accordance with the CRP. By repeating this procedure an infinite number of times, we obtain a path in a tree for the customer. Here, each restaurant corresponds to a level of the tree structure. For all customers, we obtain a tree of infinite breadth, which we call an infinite tree.

3.3 Hierarchical Latent Dirichlet Allocation

Hierarchical latent Dirichlet allocation (hLDA) [4] is a hierarchical topic model for *unimodal* data. hLDA generates an unbounded tree-structured hierarchy of latent topics. The tree is generated from nCRP, in which each document is considered as a customer and each latent topic is considered as a table (or restaurant), as illustrated in **Fig. 1**. The generative process of hLDA is described below, where we assume symmetric Dirichlet hyperparameters, following [4]:

(1) For each topic $t \in \mathcal{T}$,

  (a) Draw a multinomial over words, $\boldsymbol{\phi}_t \sim \text{Dirichlet}(\beta)$.

(2) For each document $d \in \{1, \ldots, D\}$,

  (a) Set node $c_{d,1}$ to the root of $\mathcal{T}$.

  (b) For each level $\ell \in \{2, \ldots, L\}$,

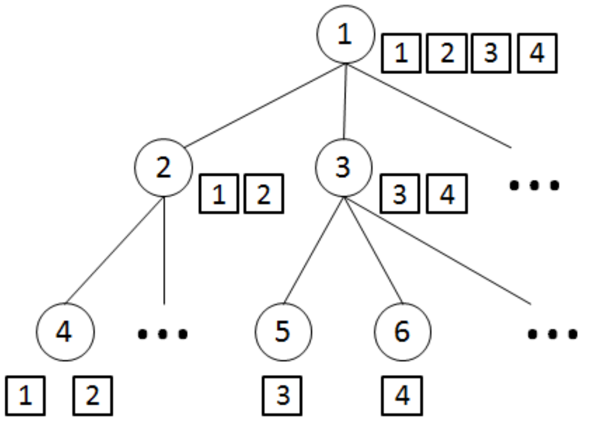


Fig. 1 An example tree generated by nCRP. Encircled numbers denote table (or topic) IDs, and boxed numbers denote customer (or document) IDs. Each customer (document) is assigned to a single path started from the root node. In this example, customer (document) 1 is assigned to tables (topics) 1, 2, and 4, while customer (document) 3 is assigned to tables (topics) 1, 3, and 5.

(i) Draw a node $c_{d,\ell}$ in accordance with Eq. (1).

(c) Draw a multinomial over levels in the tree, $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$.

(d) For each word $w_{d,n}$ (where $n \in \{1,\dots,N_d\}$),

(i) Draw level $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$.

(ii) Draw word $w_{d,n} \sim \text{Multinomial}(\boldsymbol{\phi}_{c_{d,z_{d,n}}})$.

Here, Step 2-(b) corresponds to nCRP described in Section 3.2 since it recursively uses CRP, as expressed in Eq. (1), along the finite number of levels. Once each document d is assigned to a path $c_{d,\cdot}$ in a tree, as illustrated in Fig. 1, a multinomial parameter over levels $\boldsymbol{\theta}_d$ is selected and then a level $z_{d,n}$ is selected for each word $w_{d,n}$ in accordance with the multinomial. Here, each level in a path specifies a node in the tree, where each node corresponds to a topic. Differently from LDA, topics in hLDA are structured in a tree hierarchy. It means that, in a tree hierarchy, the topics close to the root indicate general topics while those close to a leaf indicate specific topics.

4. h-SymCorrLDA

In the previous model [5], the model first generates topics (or nodes in a topic hierarchy) for one document mode, which we refer to as a *pivot mode*. For the other mode, it then uses the topics that were already generated in the pivot mode. However, this kind of model has the disadvantage that it requires a pivot mode to be specified in advance. On the other hand, our h-SymCorrLDA incorporates a hidden variable to control the pivot mode, which is similar to that of the non-hierarchical SymCorrLDA [3]. Our model generates a flag that specifies a pivot mode for each word, adjusting the probability of being a pivot in each document mode in accordance with a multinomial distribution. In other words, from the data, h-SymCorrLDA estimates the best pivot mode at the word level in each document. The pivot flag $x_i^{(k)} = k$ for an arbitrary mode k indicates that the pivot mode for the word $w_i^{(k)}$ is its own mode k, and $x_i^{(k)} = j$ indicates that the pivot mode for $w_i^{(k)}$ is another mode j, which is different from its own mode k. h-SymCorrLDA's generative process is described as follows.

(1) For each topic $t \in \mathcal{T}$,

(a) Draw a multinomial over words, $\boldsymbol{\phi}_t^{(k)} \sim \text{Dirichlet}(\beta)$.

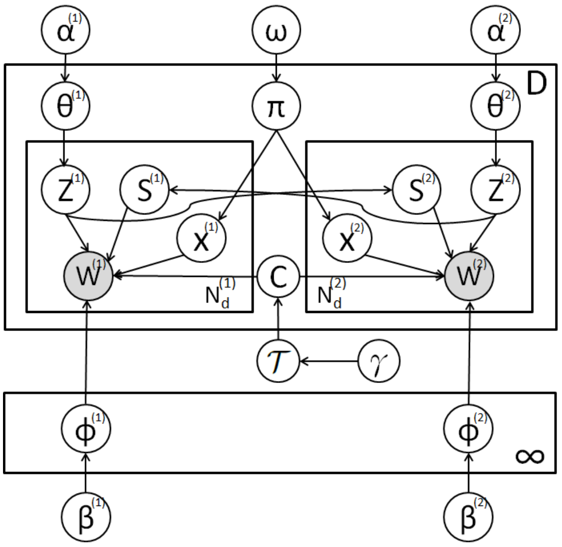


Fig. 2 Graphical model of h-SymCorrLDA.

(2) For each document $d \in \{1,\dots,D\}$,

(a) Draw a multinomial over pivot flags, $\boldsymbol{\pi}_d \sim \text{Dirichlet}(\omega)$.

(b) Set node $c_{d,1}$ to the root of $\mathcal{T}$.

(c) For each level $\ell \in \{2,\dots,L\}$,

(i) Draw a node $c_{d,\ell}$ in accordance with Eq. (1).

(d) For each mode $k \in \{1,\dots,K\}$,

(i) Draw a multinomial over levels in the tree, $\boldsymbol{\theta}_d^{(k)} \sim \text{Dirichlet}(\alpha^{(k)})$.

(ii) For each word $w_{d,n}^{(k)}$ (where $n \in \left\{1,\dots,N_d^{(k)}\right\}$),

(A) Draw a pivot flag $x_{d,n}^{(k)} \sim \text{Multinomial}(\boldsymbol{\pi}_d)$.

(B) If $x_{d,n}^{(k)} = k$, draw level $z_{d,n}^{(k)} \sim \text{Multinomial}(\boldsymbol{\theta}_d^{(k)})$.

(C) If $x_{d,n}^{(k)} = j \neq k$, draw level $s_{d,n}^{(k)} \sim \text{Uniform}(z_1^{(j)},\dots,z_{N_d^{(j)}}^{(j)})$

(D) Draw word $w_{d,n}^{(k)} \sim \text{Multinomial}(\delta_{x_{d,n}^{(k)}=k}\boldsymbol{\phi}_{z_{d,n}^{(k)}}^{(k)} + (1-\delta_{x_{d,n}^{(k)}=k})\boldsymbol{\phi}_{s_{d,n}^{(k)}}^{(k)})$.

Here, the indicator function δ takes the value 1 when the designated event occurs and 0 otherwise. Step 2-(c) corresponds to the nCRP since it recursively uses the CRP with Eq. (1) along the finite number of levels.

Figure 2 shows a graphical model representation of h-SymCorrLDA. For reference, **Fig. 3** corresponds to Li et al.'s model [5], which is referred to as h-CorrLDA in this paper. As you can see in these figures, h-CorrLDA can be regarded as a special case of h-SymCorrLDA when the pivot mode is specified in advance.

Suppose the number of modes is two ($k \in \{1, 2\}$), h-SymCorrLDA's full conditional distribution for sampling a path is given by:

$$p(\mathbf{c}_d|\mathbf{w},\mathbf{z},\mathbf{s},\mathbf{x},\mathbf{c}_{-d},\gamma,\boldsymbol{\beta},\omega)$$
$$\propto p(\mathbf{w}_d^{(1)},\mathbf{w}_d^{(2)}|\mathbf{w}_{-d}^{(1)},\mathbf{w}_{-d}^{(2)},\mathbf{z},\mathbf{s},\mathbf{x},\mathbf{c},\boldsymbol{\beta},\omega)p(\mathbf{c}_d|\mathbf{c}_{-d},\gamma)$$

where $\mathbf{w} = \{\mathbf{w}^{(1)},\mathbf{w}^{(2)}\}$, $\mathbf{z} = \{\mathbf{z}^{(1)},\mathbf{z}^{(2)}\}$, $\mathbf{s} = \{\mathbf{s}^{(1)},\mathbf{s}^{(2)}\}$, and $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)}\}$. Here, $\mathbf{w}^{(k)} = \{w_{d,n}^{(k)}\}$, $\mathbf{z}^{(k)} = \{z_{d,n}^{(k)}\}$, and $\mathbf{s}^{(k)} = \{s_{d,n}^{(k)}\}$ where $k \in \{1,\dots,K\}$, $d \in \{1,\dots,D\}$, and $n \in \left\{1,\dots,N_d^{(k)}\right\}$. The second term on the right-hand side indicates a prior given by the

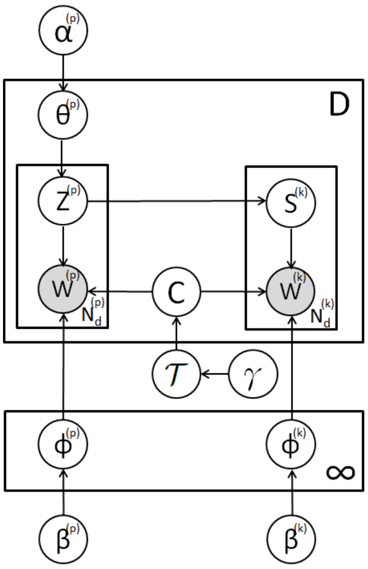


Fig. 3 Graphical model of h-CorrLDA.

nCRP that was mentioned in Section 3.2. The first term is given by:

$$p(\mathbf{w}_d^{(1)}, \mathbf{w}_d^{(2)} | \mathbf{w}_{-d}^{(1)}, \mathbf{w}_{-d}^{(2)}, \mathbf{z}, \mathbf{s}, \mathbf{x}, \mathbf{c}, \boldsymbol{\beta}, \omega)$$

$$= \prod_{\ell=1}^{L} \frac{\Gamma(n_{c_{-d,\ell},\cdot}^{(1)} + W^{(1)}\beta^{(1)})}{\prod_{w^{(1)}} \Gamma(n_{c_{-d,\ell},w^{(1)}}^{(1)} + \beta^{(1)})} \cdot \frac{\prod_{w^{(1)}} \Gamma(n_{c_{-d,\ell},w^{(1)}}^{(1)} + n_{c_{-d,\ell},w^{(1)}}^{(1)} + \beta^{(1)})}{\Gamma(n_{c_{-d,\ell},\cdot}^{(1)} + n_{c_{-d,\ell},\cdot}^{(1)} + W^{(1)}\beta^{(1)})}$$

$$\cdot \frac{\Gamma(n_{c_{-d,\ell},\cdot}^{(2)} + W^{(2)}\beta^{(2)})}{\prod_{w^{(2)}} \Gamma(n_{c_{-d,\ell},w^{(2)}}^{(2)} + \beta^{(2)})} \cdot \frac{\prod_{w^{(2)}} \Gamma(n_{c_{-d,\ell},w^{(2)}}^{(2)} + n_{c_{-d,\ell},w^{(2)}}^{(2)} + \beta^{(2)})}{\Gamma(n_{c_{-d,\ell},\cdot}^{(2)} + n_{c_{-d,\ell},\cdot}^{(2)} + W^{(2)}\beta^{(2)})}$$

where superscript ‘(k)’ indicates that the variable is for mode k. In addition, each word’s topic level is sampled as either $\mathbf{z}$ or $\mathbf{s}$; therefore, count variables ‘n’ are summed over both $\mathbf{z}$ and $\mathbf{s}$.

A full conditional distribution for collapsed Gibbs sampling of this model is given by the following equations.

$$p(z_{d,n}^{(k)} = \ell, x_{d,n}^{(k)} = k | \mathbf{w}^{(k)}, \mathbf{z}_{-(d,n)}^{(k)}, \mathbf{x}_{-(d,n)}, \mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega)$$

$$\propto \frac{m_{d,k}^{-(d,n)} + \omega}{m_{d,k}^{-(d,n)} + \sum_{k' \neq k} m_{d,k'} + K\omega} \cdot \frac{n_{d,\ell}^{(k),-(d,n)} + \alpha^{(k)}}{n_{d,\cdot}^{(k),-(d,n)} + L\alpha^{(k)}}$$

$$\cdot \frac{n_{c_{d,\ell},w^{(k)}}^{(k),-(d,n)} + \beta^{(k)}}{n_{c_{d,\ell},\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}} \tag{2}$$

$$p(s_{d,n}^{(k)} = \ell, x_{d,n}^{(k)} = j | \mathbf{w}^{(k)}, \mathbf{z}^{(j)}, \mathbf{z}^{(k)}, \mathbf{s}_{-(d,n)}^{(k)}, \mathbf{x}_{-(d,n)}, \mathbf{c}, \beta^{(k)}, \omega)$$

$$\propto \frac{m_{d,j}^{-(d,n)} + \omega}{m_{d,j}^{-(d,n)} + \sum_{j' \neq j} m_{d,j'} + K\omega} \cdot \frac{n_{d,\ell}^{(j)}}{N_d^{(j)}} \cdot \frac{n_{c_{d,\ell},w^{(k)}}^{(k),-(d,n)} + \beta^{(k)}}{n_{c_{d,\ell},\cdot}^{(k),-(d,n)} + W^{(k)}\beta^{(k)}} \tag{3}$$

A brief derivation of the full conditional distribution is given in Appendix. For example, suppose that we have text-annotated images where modes k and j are image mode and text mode, respectively. As you can see in the first terms of the right-hand side of Eqs. (2) and (3), the more often the pivot flag for each visual-word token in image mode k of document d specifies the same mode k, the more likely one of the image mode’s topics is selected or a new topic is generated. On the other hand, the more often the pivot flag for each visual-word token in image mode k of document d specifies text mode j, the more likely one of the text mode’s topics is selected and therefore the text annotations help the better topic assignments, as in the situation with “dog” and “puppy” that we mentioned in the introduction.

5. Experiments

In this section, we show the effectiveness of our h-SymCorrLDA by comparing our model with some baseline models in terms of two evaluation metrics. We briefly describe the datasets that we used for experiments in Section 5.1 and some experimental settings in Section 5.2. We then describe three metrics that we used for evaluation in Section 5.3. In Section 5.4, we give the evaluation results and show that our model performs more effectively than the baseline models.

5.1 Datasets

For experiments, we used MIRFLICKR-25000 [9], its subset that we call MIRFLICKR-small, and UIUC-Sports datasets [10]. MIRFLICKR-25000 consists of 25,000 images that were sampled from Flickr, where each image is annotated by a user. The average number of annotated tags per image is 8.94. Each image is also associated with class labels. The class labels are structured as a tree, and the number of classes is 24. For MIRFLICKR-small, we selected 100 images per class, and the resulting number of images is 2382 since only 82 images are associated with class *baby*. UIUC-Sports consists of 8 classes about sports. We used the dataset used by Zheng et al. [11], where the number of images per class ranges from 137 (*bocce*) to 330 (*croquet*), and the total number of images is 1,792.

5.2 Experimental Settings

For MIRFLICKR-25000 and MIRFLICKR-small, we removed the tags that appear less than 20 times, resulting in 1,386 types of tags for either case. Moreover, we removed images associated with no classes. For images associated with multiple classes, we selected a single class in the following manner.

- For each image, select the classes that are positioned at the deepest node in the class hierarchy among multiple classes associated with each image.
- If the image is associated with multiple same-depth classes, select the class that appears least frequently in the dataset.

We extracted SIFT descriptors [12] using dense sampling [6] for the MIRFLICKR-2500 and MIRFLICKR-small datasets. We set the grid size to 30×30 pixels and set the scale to 30 pixels. We then extracted visual words using a k-means algorithm, resulting in 1,000 visual word types. For the UIUC-Sports dataset, we used the same settings as used by Zheng et al. [11]. We set the grid size to 8×8 pixels and set the scale to 16 pixels. The number of visual word types is 240. We give a summary of the three datasets in **Table 2**.

To carry out experiments, we randomly selected 20% of images from each dataset to obtain a *test set*. Using the remaining 80% of images, we performed four-fold cross validation where we used four combinations of a *training set* and a *validation set*.

For all the hierarchical topic models, we set the maximum level of the topic tree as $L = 4$. We assumed the symmetric Dirichlet hyperparameter $\alpha = 0.1$ for all the models. We also assumed the

Table 2 Summary of datasets we used.

	MIRFLICKR-small		MIRFLICKR-25000		UIUC-Sports	
	visual words	tags	visual words	tags	visual words	tags
# images	2,382		25,000		1,792	
# word tokens	511,818	27,774	5,379,098	94,282	2,996,832	25,618
# word types	1,000	1,386	1,000	1,386	240	161

symmetric Dirichlet hyperparameter $\omega = 1.0$, in accordance with Fukumasu et al. [3]. We set the nCRP parameter $\gamma = 1.0$, in accordance with Blei et al. [4]. As for β, we assumed $\beta^{(k)} = \frac{\lambda}{W^{(k)}}$ for each mode k, where λ is a parameter. This enables consideration of the difference in the number of word types between visual words and tags. For LDA and hLDA, we mixed visual words and tags to make up a single-mode representation. We used collapsed Gibbs sampling [13] for estimating the models. We terminated the Gibbs sampling procedure when the percentage change of test-set log-likelihood (as defined in Section 5.3) is less than 0.1%.

5.3 Evaluation Metrics

We use two metrics to evaluate the models: one is a test-set log-likelihood and the other is normalized mutual information (NMI). The test-set log-likelihood is a metric for measuring the generalization ability while NMI is a metric for measuring the clustering ability.

The test-set log-likelihood is given by:

$$LL(\mathbf{w}^{test}) = \frac{\sum_d \sum_n \log P(w_{d,n}^{test}|\mathbf{w}^{train}, \mathcal{M})}{\sum_d N_d^{test}}$$

where $\mathbf{w}^{test}$ and $\mathbf{w}^{train}$ denote the test set and the training set, respectively. N_d^{test} denotes the number of words in document d in the test set. $\mathcal{M}$ indicates the model to be evaluated. It is generally believed that the higher the test-set log-likelihood, the better the generalization ability of the model. For this experiment, we sampled 20% of the tags from each image as the test set, and the remaining tags and all the visual words are used as the training set for estimating the models.

NMI is widely used for the evaluation of the clustering ability. NMI between a set of classes **a** and a set of clusters **b** is given by:

$$NMI(\mathbf{a}, \mathbf{b}) = \frac{MI(\mathbf{a}, \mathbf{b})}{\{H(\mathbf{a}) + H(\mathbf{b})\}/2}$$

where $MI(\mathbf{a}, \mathbf{b})$ is the mutual information between the classes and the clusters:

$$MI(\mathbf{a}, \mathbf{b}) = \sum_{a_i \in \mathbf{a}} \sum_{b_j \in \mathbf{b}} P(a_i, b_j) \mathbf{log} \frac{P(a_i, b_j)}{P(a_i)P(b_j)}$$

$H(\mathbf{a})$ and $H(\mathbf{b})$ are the marginal entropies of the classes and the clusters, respectively:

$$H(\mathbf{a}) = -\sum_{a_i \in \mathbf{a}} P(a_i) \mathbf{log} P(a_i)$$

$$H(\mathbf{b}) = -\sum_{b_j \in \mathbf{b}} P(b_j) \mathbf{log} P(b_j)$$

NMI takes a value from 0 to 1. The higher the NMI is, the more the dependency lies between the classes and the clusters and therefore the greater the clustering ability is. For a hierarchical clustering of annotated images, we sampled a level that each annotated image d is associated with, in accordance with the per-image multinomial over levels $\boldsymbol{\theta}_d^{(\cdot)}$. Each annotated image is associated with a path in the topic tree, so the sampled level specifies a node in the tree. As a result, the hierarchical clustering of annotated images is achieved.

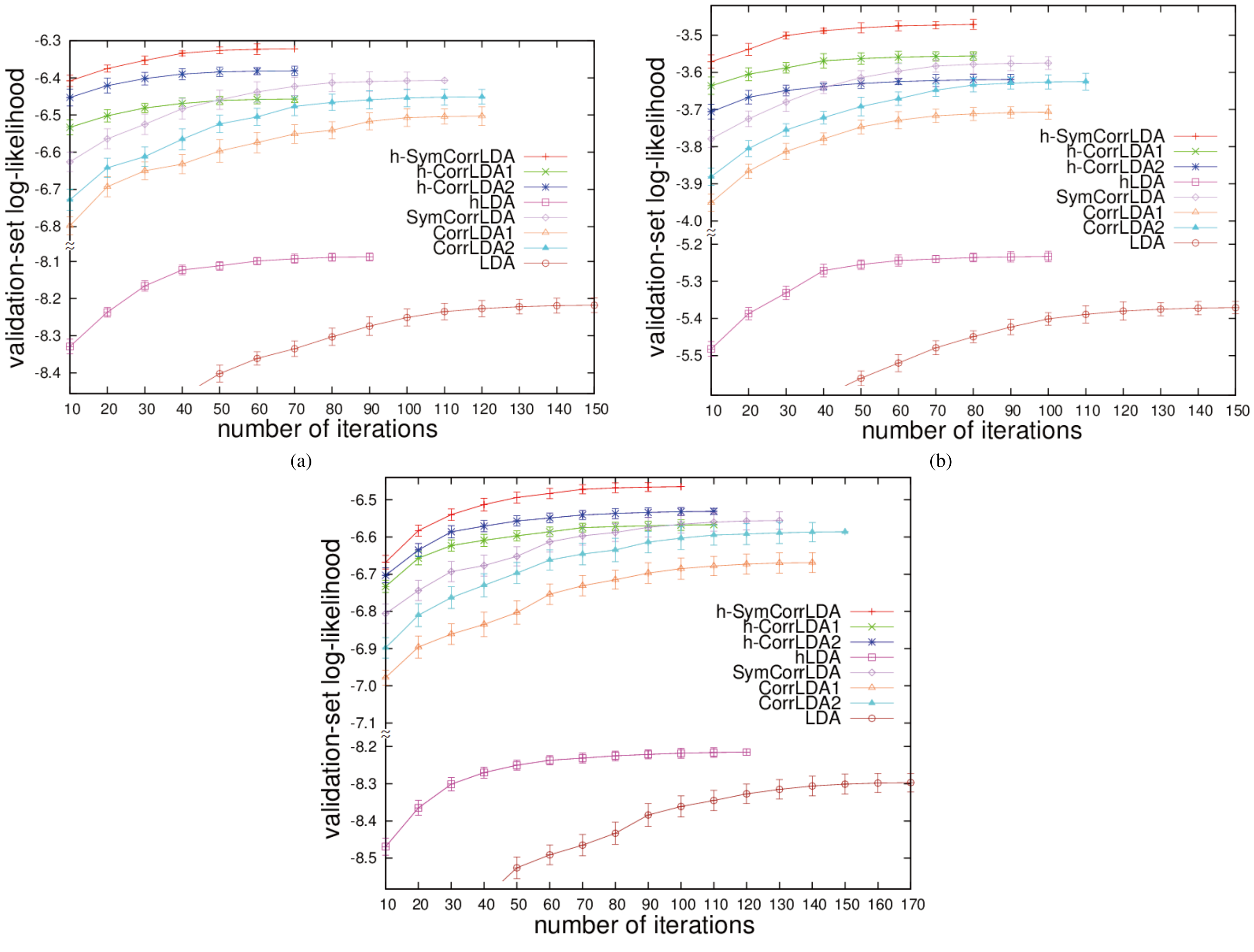
Table 3 Test-set log-likelihood with MIRFLICKR-small, UIUC-Sports, and MIRFLICKR-25000 datasets.

	MIRFLICKR-small	MIRFLICKR-25000	UIUC-Sports
h-SymCorrLDA	**-6.315**	**-6.453**	**-3.465**
h-CorrLDA1	-6.449	-6.560	-3.553
h-CorrLDA2	-6.373	-6.525	-3.613
hLDA	-8.082	-8.207	-5.229
SymCorrLDA	-6.406	-6.553	-3.561
CorrLDA1	-6.501	-6.665	-3.701
CorrLDA2	-6.445	-6.583	-3.618
LDA	-8.203	-8.289	-5.352

5.4 Results

Figure 4 shows the results of validation-set log-likelihood for three datasets: MIRFLICKR-small, MIRFLICKR-25000, and UIUC-Sports. h-CorrLDA1 and h-CorrLDA2 indicate hierarchical CorrLDA (h-CorrLDA) models, as shown in Fig. 3, when visual words are specified as the pivot mode and when tags are specified as the pivot mode, respectively. CorrLDA1 and CorrLDA2 indicate the Correspondence topic models (CorrLDA) that we mentioned in Section 2 when visual words are specified as the pivot mode and when tags are specified as the pivot mode, respectively. For all the non-hierarchical topics models: SymCorrLDA, CorrLDA, and LDA, we set the number of topics to $\{200, 400, 600, 800\}$ and then selected the optimal number of topics that brings the highest validation-set log-likelihood over four cross-validation runs. As you can see in Fig. 4, h-SymCorrLDA performs most effectively since h-SymCorrLDA considers mutual dependencies between images and tags while h-CorrLDA only considers a one-way dependency from images to tags. Moreover, we can also see that the validation-set log-likelihood with hierarchical topic models was higher than that with non-hierarchical topic models probably because hierarchical topic models assign each annotated image only to the topic nodes along a specific path in the topic hierarchy, resulting in a sparse representation in the topic space. These results were obtained when $\lambda = 1{,}000$. For the definition of λ, see Section 5.2. We also experimented by varying λ; however, we confirmed that the results were not affected by λ.

We finally obtained the log-likelihood of the test set that was not used for the four-fold cross validation in the previous experiments. We used the models estimated with both the training set and the validation set. For all the non-hierarchical topics models:

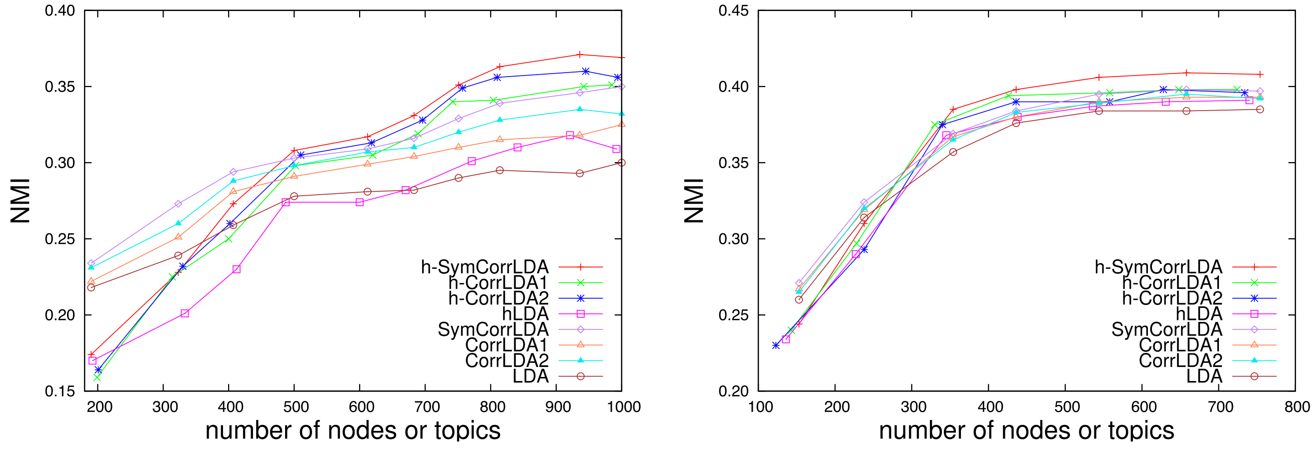


(a)

(b)

(c)

Fig. 4 Validation-set log-likelihood with (a) MIRFLICKR-small dataset, (b) UIUC-Sports dataset, and (c) MIRFLICKR-25000 dataset. Horizontal axis denotes number of iterations of Gibbs sampling. Results were averaged over four cross-validation runs. Error bars represent one sample standard deviation.



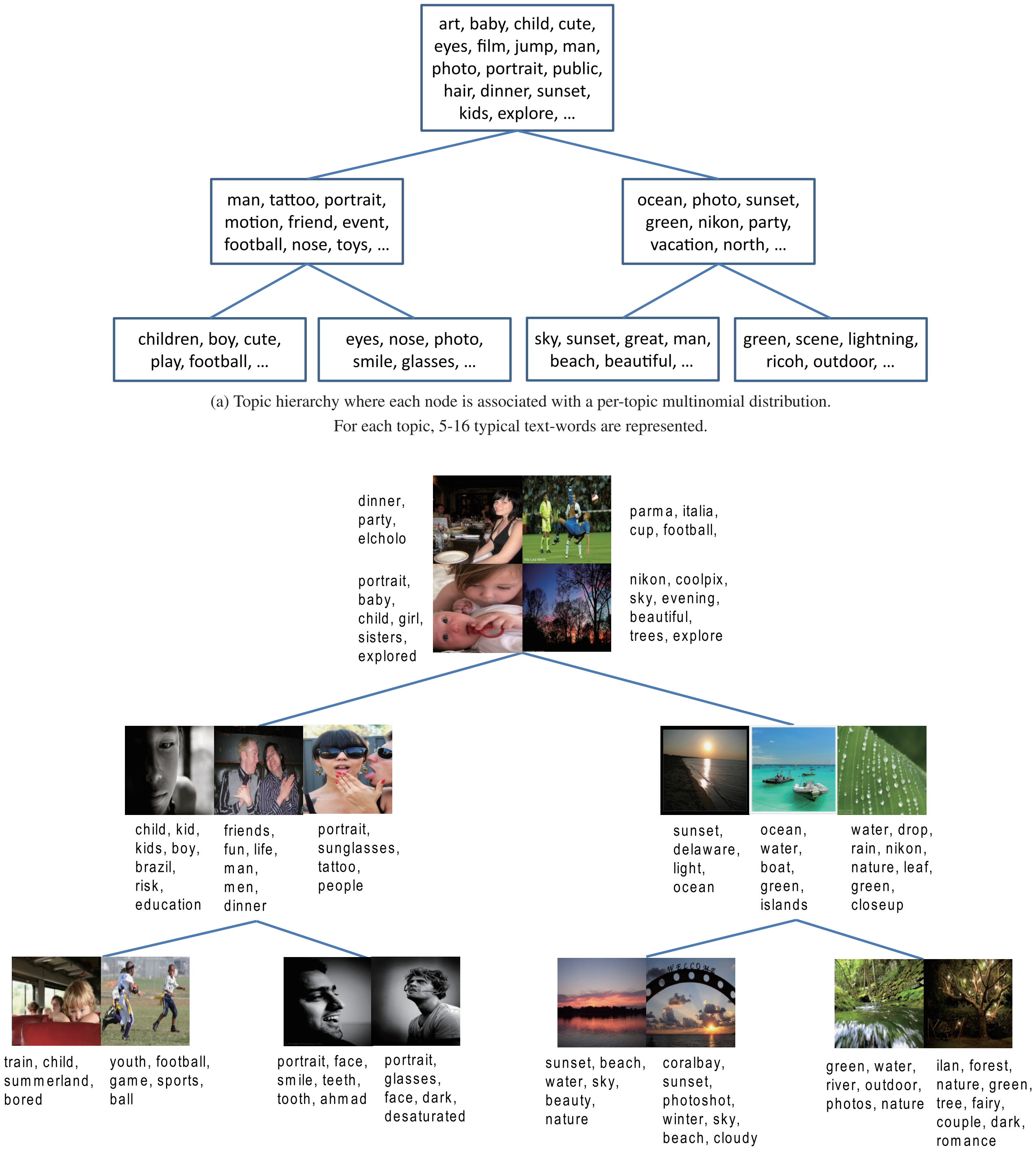
(a)

(b)

Fig. 5 NMI with (a) MIRFLICKR-small dataset and (b) UIUC-Sports dataset. For hierarchical topic models, the estimated number of nodes depends on parameter λ. The larger λ is, the fewer nodes are generated. We experimented varying λ and then plotted the estimated number of nodes and the corresponding NMI.

SymCorrLDA, CorrLDA, and LDA, we used the optimal number of topics that was selected via the four-fold cross validation. From the results shown in **Table 3**, h-SymCorrLDA achieved the highest test-set log-likelihood, indicating the most powerful prediction ability.

We conducted the experiments on NMI with MIRFLICKR-


art, baby, child, cute, eyes, film, jump, man, photo, portrait, public, hair, dinner, sunset, kids, explore, …
man, tattoo, portrait, motion, friend, event, football, nose, toys, …
ocean, photo, sunset, green, nikon, party, vacation, north, …
children, boy, cute, play, football, …
eyes, nose, photo, smile, glasses, …
sky, sunset, great, man, beach, beautiful, …
green, scene, lightning, ricoh, outdoor, …
dinner, party, elcholo
parma, italia, cup, football,
portrait, baby, child, girl, sisters, explored
nikon, coolpix, sky, evening, beautiful, trees, explore
child, kid, kids, boy, brazil, risk, education
friends, fun, life, man, men, dinner
portrait, sunglasses, tattoo, people
sunset, delaware, light, ocean
ocean, water, boat, green, islands
water, drop, rain, nikon, nature, leaf, green, closeup
train, child, summerland, bored
youth, football, game, sports, ball
portrait, face, smile, teeth, tooth, ahmad
portrait, glasses, face, dark, desaturated
sunset, beach, water, sky, beauty, nature
coralbay, sunset, photoshot, winter, sky, beach, cloudy
green, water, river, outdoor, photos, nature
ilan, forest, nature, green, tree, fairy, couple, dark, romance

(a) Topic hierarchy where each node is associated with a per-topic multinomial distribution. For each topic, 5-16 typical text-words are represented.

(b) Alternative representation of the same topic hierarchy where each node is associated with typical images. For each image, the corresponding text annotations are also represented.

Fig. 6 An example of topic hierarchy estimated using h-SymCorrLDA with MIRFLICKR-small.

small and UIUC-Sports, omitting that with MIRFLICKR-25000 for simplicity. To obtain NMI, we used the models estimated with both the training set and the validation set. **Figure 5** shows the result of NMI. Each model demonstrated the highest ability where the number of topics (or nodes) is 500-900 for the MIRFLICKR-small and 400-600 for the UIUC-Sports. Under these conditions, h-SymCorrLDA outperforms the other models in terms of NMI, as shown in Fig. 5. The difference from h-CorrLDA2 in Fig. 5 (a) or h-CorrLDA1 in Fig. 5 (b) is not very large; however, those models require a pivot mode to be specified in advance while h-SymCorrLDA has the advantage not to require such previous knowledge.

We give an example of topic hierarchy estimated using h-SymCorrLDA with MIRFLICKR-small, in **Fig. 6**. As can be seen in this example, the right-hand side topics represent natural scenic views while the left-hand side topics represent portrait photos. Moreover, the topics corresponding to the upper nodes are more general while those corresponding to the lower nodes are more specific.

6. Conclusions

In this paper, we proposed a new hierarchical topic model, h-SymCorrLDA, which can be applied to multimodal data. This model enabled the discovery of latent topic hierarchies considering mutual dependencies among modes, which previous models did not address. We demonstrated that our model outperformed several baselines in terms of test-set log-likelihood and normalized mutual information through experiments with three annotated image datasets. More detailed evaluation is left for future work.

Our h-SymCorrLDA can be applied to other multimodal data, such as multilingual comparable documents and video data. Experiments of such applications are left for our future work. Similarly to hLDA, h-SymcorrLDA assumes each data is assigned to only a single path in the latent topic hierarchy. We plan to extend our model so that each data is allowed to be associated with a mixture of paths, as in a recent work of Ref. [14].

Acknowledgments This work was supported in part by the Grant-in-Aid for Scientific Research (#23300039 and #15H02703) from JSPS, Japan, and in part by a joint research project of the National Institute of Informatics and Kobe University.

References

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[2] Blei, D.M. and Jordan, M.I.: Modeling annotated data, *Proc. 26th Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, pp.127–134 (2003).

[3] Fukumasu, K., Eguchi, K. and Xing, E.: Symmetric Correspondence Topic Models for Multilingual Text Analysis, *Advances in Neural Information Processing Systems*, Vol.25, pp.1295–1303 (2012).

[4] Blei, D.M., Griffiths, T., Jordan, M. and Tenenbaum, J.: Hierarchical topic models and the nested Chinese restaurant process, *Advances in neural Information Processing Systems*, Vol.16, pp.106–114 (2004).

[5] Li, L.-J., Wang, C., Lim, Y., Blei, D.M. and Fei-Fei, L.: Building and using a semantivisual image hierarchy, *Proc. 2010 IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, California, USA, pp.3336–3343 (2010).

[6] Fei-Fei, L. and Perona, P.: A bayesian hierarchical model for learning natural scene categories, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, Vol.2, pp.524–531, IEEE (2005).

[7] Wang, C., Blei, D.M. and Li, F.-F.: Simultaneous Image Classification and Annotation, *Proce. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR 2009*), Miami, Florida, USA, pp.1903–1910 (2009).

[8] Aldous, D.: Exchangeability and Related Topics, *École d'Été de probabilités de Saint-Flour XIII–1983*, Springer, Berlin, pp.1–198 (1985).

[9] Huiskes, M.J. and Lew, M.S.: The MIR Flickr Retrieval Evaluation, *Proc. 2008 ACM International Conference on Multimedia Information Retrieval*, Vancouver, Canada, pp.39–43 (2008).

[10] Li, L.-J. and Fei-Fei, L.: What, where and who? Classifying events by scene and object recognition, *IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp.1–8 (2007).

[11] Zheng, Y., Zhang, Y.-J. and Larochelle, H.: Topic Modeling of Multimodal Data: an Autoregressive Approach, *Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, USA, pp.1370–1377 (2014).

[12] Lowe, D.G.: Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).

[13] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. the National Academy of Sciences of the United States of America*, Vol.101, No.Suppl 1, pp.5228–5235 (2004).

[14] Ahmed, A., Hong, L. and Smola, A.: Nested chinese restaurant franchise process: Applications to user tracking and document modeling, *Proc. 30th International Conference on Machine Learning* (*ICML-13*), pp.1426–1434 (2013).

Appendix

We give here a brief derivation of the full conditional distribution shown in Eq. (2).

$$p(z^{(k)}_{d,n} = \ell, x^{(k)}_{d,n} = k|\mathbf{w}^{(k)}, \mathbf{z}^{(k)}_{-(d,n)}, \mathbf{x}_{-(d,n)}, \mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega)$$

$$= \frac{p(\mathbf{w}^{(k)}, \mathbf{z}^{(k)}, \mathbf{x}|\mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega)}{p(\mathbf{w}^{(k)}, \mathbf{z}^{(k)}_{-(d,n)}, \mathbf{x}_{-(d,n)}|\mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega)}$$

$$\propto \frac{p(\mathbf{w}^{(k)}, \mathbf{z}^{(k)}, \mathbf{x}|\mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega)}{p(\mathbf{w}^{(k)}_{-(d,n)}, \mathbf{z}^{(k)}_{-(d,n)}, \mathbf{x}_{-(d,n)}|\mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega)}$$

$$= \frac{\int p(\mathbf{w}^{(k)}, \mathbf{z}^{(k)}, \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega) d\boldsymbol{\theta} d\boldsymbol{\phi} d\boldsymbol{\pi}}{\int p(\mathbf{w}^{(k)}_{-(d,n)}, \mathbf{z}^{(k)}_{-(d,n)}, \mathbf{x}_{-(d,n)}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}|\mathbf{c}, \alpha^{(k)}, \beta^{(k)}, \omega) d\boldsymbol{\theta} d\boldsymbol{\phi} d\boldsymbol{\pi}}$$

$$= \frac{\begin{array}{c}\int p(\mathbf{x}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\omega) d\boldsymbol{\pi} \int p(\mathbf{z}^{(k)}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta} \\ \times \int p(\mathbf{w}^{(k)}|\mathbf{z}^{(k)}, \mathbf{c}, \mathbf{x}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\beta^{(k)}) d\boldsymbol{\phi}\end{array}}{\begin{array}{c}\int p(\mathbf{x}_{-(d,n)}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\omega) d\boldsymbol{\pi} \int p(\mathbf{z}^{(k)}_{-(d,n)}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta} \\ \times \int p(\mathbf{w}^{(k)}_{-(d,n)}|\mathbf{z}^{(k)}_{-(d,n)}, \mathbf{c}, \mathbf{x}_{-(d,n)}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\beta^{(k)}) d\boldsymbol{\phi}\end{array}}$$

$$= \frac{m^{-(d,n)}_{d,k} + \omega}{m^{-(d,n)}_{d,k} + \sum_{k' \neq k} m_{d,k'} + K\omega} \cdot \frac{n^{(k),-(d,n)}_{d,\ell} + \alpha^{(k)}}{n^{(k),-(d,n)}_{d,\cdot} + L\alpha^{(k)}} \cdot \frac{n^{(k),-(d,n)}_{c_{d,\ell}, w^{(k)}} + \beta^{(k)}}{n^{(k),-(d,n)}_{c_{d,\ell},\cdot} + W^{(k)}\beta^{(k)}}$$

Equation (3) can be easily derived in the same way.

Editor's Recommendation

IPSJ Kansai Branch selected excellent papers out of the papers presented at the IPSJ Kansai-Branch Convention 2014 for recommendation to the Journal of Information Processing. The session chairpersons and committee members of the convention nominated four candidates from the 23 long papers presented at the convention, and then each of the candidates was reviewed by two referees. After careful discussion among the committee members, we finally selected two papers to be recommended.

This paper proposes a Bayesian model to discover latent hierarchies from multimodal data, and demonstrates its effectiveness through experiments with annotated image datasets. We found significant novelty of the proposed model in this research area. Therefore, we have decided to recommend that this paper would be submitted to the Journal of Information Processing.

(Toru Fujiwara, Chairman of IPSJ Kansai Branch)

Takuji Shimamawari is currently pursuing a masters degree at the Graduate School of System Informatics, Kobe University, Japan.

Koji Eguchi is an associate professor at the Graduate School of System Informatics, Kobe University, Japan. His research interests include information retrieval, statistical machine learning, and data mining.

Atsuhiro Takasu is a professor at the National Institute of Informatics, Japan. His research interests are database systems and machine learning. He is a member of ACM, IEEE, IPSJ, DBSJ and JSAI.