Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Hierarchical topic modeling with nested hierarchical Dirichlet process*

Yi-qun DING[1], Shan-ping LI[†‡1], Zhen ZHANG[1], Bin SHEN[2]

([1]*School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)

([2]*State Street Hangzhou, Hangzhou 310000, China*)

[†]E-mail: shan@zju.edu.cn

Received Nov. 15, 2008; Revision accepted Apr. 10, 2009; Crosschecked Apr. 29, 2009

Abstract: This paper deals with the statistical modeling of latent topic hierarchies in text corpora. The height of the topic tree is assumed as fixed, while the number of topics on each level as unknown a priori and to be inferred from data. Taking a nonparametric Bayesian approach to this problem, we propose a new probabilistic generative model based on the nested hierarchical Dirichlet process (nHDP) and present a Markov chain Monte Carlo sampling algorithm for the inference of the topic tree structure as well as the word distribution of each topic and topic distribution of each document. Our theoretical analysis and experiment results show that this model can produce a more compact hierarchical topic structure and captures more fine-grained topic relationships compared to the hierarchical latent Dirichlet allocation model.

Key words: Topic modeling, Natural language processing, Chinese restaurant process, Hierarchical Dirichlet process, Markov chain Monte Carlo, Nonparametric Bayesian statistics

doi:10.1631/jzus.A0820796 **Document code:** A **CLC number:** O212.8; H03

INTRODUCTION

Given how fast documents are generated in large corporations and over the Internet, there is an increasing need for automatic document analysis technologies, such as cluster-based information retrieval, organization of Internet search results, document classification, and document vetting (e.g., detection of similar, low content-bearing documents). Topic modeling, as one answer to these needs, automatically extracts common word usage patterns (or 'topics'), given a group of documents, by using probabilistic generative models. Representing documents in the low-dimensional topic space rather than the high-dimensional word space, which results in a dramatic reduction of the original data, is the key to both automatic and manual processing of large document collections (Wei and Croft, 2006; Walker and Ringger, 2008; Zhang *et al*., 2008).

One problem with some topic models is the correlations among the random variables, which can easily make inference algorithms like Gibbs sampling get trapped in local maxima. The resulting topic structure often has many duplicated or similar topics, failing to capture all the semantic relationships among the topics.

We propose a hierarchical generative model for modeling the hierarchical topic structure in text corpora. The specially designed relationship among random variables in our model helps sampling algorithms escape local maxima. The proposed model can produce a more compact topic structure and discover more fine-grained topic relationships while claiming similar predictive capacities compared to current topic models.

‡ Corresponding author

* Project (No. 60773180) supported by the National Natural Science Foundation of China

RELATED WORK

The research of learning latent topics and topic relationship embedded in a corpus of documents has a longer history than probabilistic topic modeling in information retrieval. Early work that borrowed a lot from existing clustering algorithms is often called 'document clustering'. Distance-based document clustering algorithms (Boley, 1998; Dhillon and Modha, 2001) optimize some objective functions of document distances based on different document representations, such as the vector space model and the suffix tree. The wide use of Euclidian distance often makes these algorithms suffer from the curse of dimensionality. Model-based document clustering algorithms (Strehl *et al*., 2000; Elkan, 2006) assume a usually simple generative model for the documents, such as a mixture of multinomial distributions, and estimate the model parameters.

Latent semantic analysis (LSA) (Deerwester *et al*., 1990) assumes that each document is a linear combination of topics (i.e., eigenvectors of the term-document matrix). The projection of documents onto the space spanned by topics can eliminate noise introduced by word choice or writing styles. The fact that the discovered topics are mutually orthogonal is not intuitive and is overly restrictive. LSA cannot deal with the polysemy problem (i.e., one word with multiple meanings), because each word has only one coordinate in the reduced space. Also, the topics are hard to interpret because they often contain negative components.

Recently probabilistic topic modeling algorithms (Blei *et al*., 2003b; Griffiths and Steyvers, 2004) have attracted a lot of research interest. These models share similar assumptions about documents with LSA: documents are mixtures of topics while topics are probability distributions over words. This idea was further developed on the basis of probabilistic generative models. Unlike LSA, probabilistic generative models do not require topics to be orthogonal to each other. Also, probabilistic generative models can cope with the polysemy problem because the same words can appear in multiple topics with a high probability. They are quite effective in discovering various topic structures, both flat (Blei *et al*., 2003b; Griffiths and Steyvers, 2004) and hierarchical (Blei *et al*., 2003a). The basic model has been extended in various ways (Rosen-Zvi *et al*., 2004; Blei and Lafferty, 2006; 2007; Li and McCallum, 2006; Wallach, 2006).

HIERARCHICAL DIRICHLET PROCESS

A brief introduction will be given here to the Dirichlet process (also known as the 'Chinese restaurant process') and the hierarchical Dirichlet process (HDP) to make this paper self-contained. The clustering property of the Dirichlet process (Blackwell and MacQueen, 1973) makes it a good mathematical tool for modeling the clusters within a group of data. φ_1, φ_2, …, φ_n are n independent and identically-distributed (i.i.d.) random variables distributed according to G. G is distributed as a Dirichlet process $DP(\alpha, G_0)$, where α is the concentration parameter and G_0 is the base distribution. Define ψ_1, ψ_2, …, ψ_K as the K distinct values (ψ_1, ψ_2, …, ψ_K are distinct with a probability of 1 if G_0 is continuous) taken on by φ_1, φ_2, …, φ_{i-1}, and let n_k be the number of $\varphi_{i'}$'s that are equal to ψ_k for $1\leq i'\leq i-1$. φ_i has the following conditional distribution with G integrated out:

$$\varphi_i|\varphi_1,\varphi_2,...,\varphi_{i-1},\alpha,G_0 \sim \sum_{k=1}^{K}\frac{n_k}{i-1+\alpha}\delta_{\psi_k}+\frac{\alpha}{i-1+\alpha}G_0, \tag{1}$$

where δ_ψ is an atom at ψ. The mixture distribution in Eq.(1) means that φ_i can be equal to a previously-seen value ψ_k with a probability proportional to n_k, or that φ_n can be equal to a new value distributed according to the base distribution G_0. We call φ_i's with the same value a cluster.

The HDP was designed to enable the sharing of clusters among multiple groups of random variables distributed according to a Dirichlet process (Teh *et al*., 2006). $\boldsymbol{\varphi}=(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_J)$ are J groups of random variables. $\boldsymbol{\varphi}_j=(\varphi_{j1}, \varphi_{j2}, \ldots, \varphi_{jn_j})$ is a group of n_j i.i.d. random variables distributed according to a distribution G_j. G_j is distributed as a Dirichlet process $DP(\alpha, G_0)$. $\psi_{j1}, \psi_{j2}, \ldots, \psi_{jK_j}$, K_j and n_{jk} are defined similarly as in Eq.(1). Given $\varphi_{j1}, \varphi_{j2}, \ldots, \varphi_{j(i-1)}$, with G_j integrated out, the conditional distribution of φ_{ji} is

$$\varphi_{ji}|\varphi_{j1},\varphi_{j2},...,\varphi_{j(i-1)},\alpha,G_0 \sim \sum_{k=1}^{K_j}\frac{n_{jk}}{i-1+\alpha}\delta_{\psi_{jk}}+\frac{\alpha}{i-1+\alpha}G_0. \tag{2}$$

In HDP, G_0 is distributed as another Dirichlet process $DP(\gamma, H)$. When φ_{ji} is drawn from the last component in the right-hand side of Eq.(2), we increase K_j by one, and φ_{ji} will be equal to the new ψ_{jK_j}, which is distributed according to G_0. With G_0 integrated out, ψ_{jK_j} has the following conditional distribution:

$$\psi_{jK_j}|\boldsymbol{\psi}_1,\boldsymbol{\psi}_2,...,\boldsymbol{\psi}_{j(K_j-1)},\gamma,H \sim \sum_{k=1}^{K}\frac{m_k}{m+\alpha}\delta_{\theta_k}+\frac{\gamma}{m+\gamma}H. \tag{3}$$

Here $\boldsymbol{\psi}_1=(\psi_{11}, \psi_{12}, \ldots, \psi_{1K_1})$, $\boldsymbol{\psi}_{j(K_j-1)}=(\psi_{j1}, \psi_{j2}, \ldots, \psi_{j(K_j-1)})$, θ_k $(1\leq k\leq K)$ are the K distinct values taken on by ψ_{jk}'s in $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_{j(K_j-1)}$, m_k is the number of the ψ_{jk}'s that are equal to θ_k, and $m=\sum m_k$. Note that with a continuous base distribution H, θ_k $(1\leq k\leq K)$ are distinct with a probability of 1. Thus, ψ_{jK_j} can take on a value shared by $\varphi_{j'i'}$'s in the same or a different group. Thus, random variables in different groups can share clusters in HDP.

NESTED HIERARCHICAL DIRICHLET PROCESS

In this section we extend the HDP to model hierarchical relationship among the topics embedded in a corpus of documents.

We are given a corpus $\boldsymbol{c}$ composed of M documents: $\boldsymbol{c}=(d_1, d_2, \ldots, d_M)$. Each document d_m $(m=1, 2, \ldots, M)$ contains a sequence of n_m words $\boldsymbol{w}_m=(w_{m1}, w_{m2}, \ldots, w_{mn_m})$, and word order in documents is considered unimportant here. The words in a document are generated by a document-specific mixture of L topics, and each topic is a multinomial distribution over words in the vocabulary. Parameters of these multinomial distributions are independently sampled from a symmetric Dirichlet distribution with parameter η. The topic mixing proportions for all documents are independently sampled from a Dirichlet distribution. A document obtains its L topics by following a path in an L-level topic tree.

Fig.1a shows a three-level topic tree in the nested hierarchical Dirichlet process (nHDP) model. To emphasize the difference with the hierarchical latent Dirichlet allocation (hLDA) model (Blei *et al*., 2003a), we present an equivalent topic tree in the hLDA model in Fig.1b.

In the hLDA model, each node in an L-level topic tree represents a topic. A document is generated by choosing a path from the root to a leaf as determined by the nest Chinese restaurant process (CRP) and by sampling the words from a document-specific mixture of the L topics on the path.

Except for the root note, nodes in an L-level nHDP model topic tree (big nodes with dashed circles in Fig.1a) represent sub-topics. Each sub-topic node is assigned to one and only one topic node (Note: the smaller shaded nodes in Fig.1a represent topics) in its parent node. Similar sub-topics are assigned to the same topic node; for example, sub-topic nodes A and A′ in Fig.1a are assigned to the same topic node contained in the root node. The root node and the other topic nodes represent topics. To produce the equivalent topic tree, sub-topics (A and A′) assigned to the same topic node are merged together (Fig.1b).

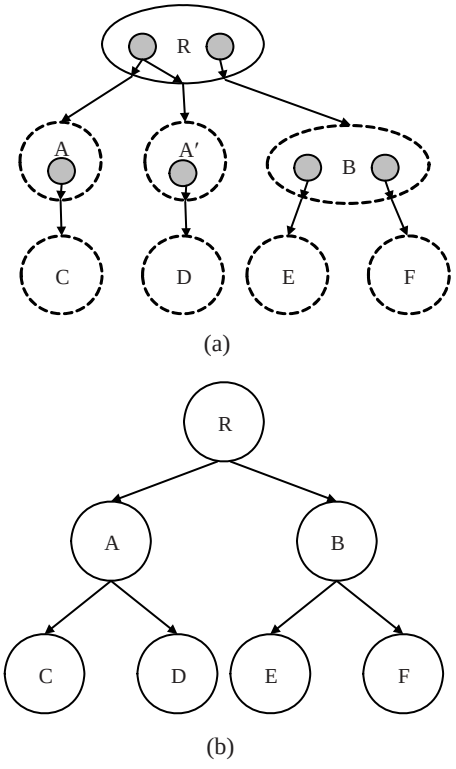

Fig.1 A three-level topic tree in nHDP (a) and its equivalent topic tree in hLDA (b)

In (a) big nodes with dashed circles represent sub-topics and the smaller shaded nodes represent topics

A path in the nHDP topic tree is composed of the root topic and $L-1$ sub-topics. A document in the nHDP model is generated by choosing a path in the topic tree and sampling the words from a document-specific mixture of the root topic and the other $L-1$ topics associated with the $L-1$ sub-topics on the path.

A topic in hLDA is a cluster of similar documents and a topic tree forms a nested clustering of the documents. A sub-topic in nHDP is a cluster of similar documents. Except for the root topic, a topic in nHDP is a cluster of sub-topics. The words for the documents assigned to a sub-topic are generated by sampling from the topic associated with the sub-topic. This two-level clustering in the nHDP model is specially designed to help sampling algorithms such as Gibbs sampling escape local maxima.

The nHDP determines the nested two-level clustering in the nHDP model. We can define nHDP by describing how a group of M documents choose their paths down the topic tree according to nHDP. Define $\boldsymbol{j}_m=(j_{m1}, j_{m2}, \ldots, j_{mL})$ to be the path of document m in the nHDP topic tree. All documents in the nHDP model share the root topic, therefore $j_{m1}=1$ ($1\leq m\leq M$). $j_{m2}, j_{m3}, \ldots, j_{mL}$ specify the $L-1$ sub-topics on the path. Given the path from the root node to the current node r on level $l-1$ as specified by $\boldsymbol{j}_{m(l-1)}=(j_{m1}, j_{m2}, \ldots, j_{m(l-1)})$, $j_{ml}=k$ ($k\geq 1$) means that document m is assigned to the kth child node (sub-topic) of node r. Let n be the number of documents that share the node r with document m. Given $\boldsymbol{j}_1, \boldsymbol{j}_2, \ldots, \boldsymbol{j}_{m-1}$ and $\boldsymbol{j}_{m(l-1)}$, j_{ml} has the following mixture distribution according to Eq.(2):

$$j_{ml}|\boldsymbol{j}_1, \boldsymbol{j}_2, \ldots, \boldsymbol{j}_{m-1}, \boldsymbol{j}_{m(l-1)}, \alpha, G_0 \sim \sum_{k=1}^{K}\frac{n_k}{n+\alpha}\delta_k+\frac{\alpha}{n+\alpha}\delta_{K+1}. \tag{4}$$

Here K is the number of child nodes of node r, and n_k is the number of documents in $(d_1, d_2, \ldots, d_{m-1})$ that are assigned to the sub-topic indexed by k. This means that document m can be assigned to one of the existing child nodes (sub-topics) with a probability proportional to n_k, or a new sub-topic (indexed by $K+1$) with a probability proportional to α.

Let S be the number of sub-topics in the topic tree, $\boldsymbol{k}=(k_1, k_2, \ldots, k_S)$ be the index of an internal topic node associated with the sub-topics, and D be the number of topic nodes in node r. When a new sub-topic is created, the index of its associated topic node k_{S+1} has the following mixture distribution according to Eq.(3):

$$k_{S+1}|\boldsymbol{k}, \gamma, H \sim \sum_{d=1}^{D}\frac{m_d}{m+\gamma}\delta_d+\frac{\gamma}{m+\gamma}\delta_{D+1}, \tag{5}$$

where m_d is the number of child nodes (sub-topics) of node r assigned to the topic node indexed by d, and $m=\sum m_d$.

Now we develop our generative document model based on nHDP. The following process describes how a corpus is generated in the nHDP model:

For each document m:

Step 1: Let j_{m1} be the root topic.

Step 2: For each level $l\in\{2, 3, \ldots, L\}$: (i) Draw a sub-topic according to Eq.(4), and set j_{ml} to be the sub-topic; (ii) If a new sub-topic is created in step (i), draw a topic according to Eq.(5), attain parameters of the multinomial distribution μ for the new topic from $Dir(\eta)$, and assign the sub-topic to it.

Step 3: Draw an L-dimensional topic mixing proportion vector $\boldsymbol{\varepsilon}_m$ from $Dir(\beta)$.

Step 4: For each word w_{mi} in document m: (i) Draw level assignment $z_{mi}\in\{1, 2, \ldots, L\}$ from $Mult(\boldsymbol{\varepsilon}_m)$. (ii) If $z_{mi}=1$, draw w_{mi} from the root topic; otherwise, from the topic to which the sub-topic on level z_{mi} of path j_m is assigned.

Hyper parameters η and β, which can be further drawn from other suitable distributions to decrease model dependency on them, are fixed in this work. While the height of the tree is fixed, the number of sub-topic nodes on each level and the number of topic nodes are not. After a group of documents are observed, the above generative process is reversed to determine the topic tree structure, the contents of the topics (distribution over words), and the distribution of documents over the topics without using any other prior knowledge.

Note the difference between the ways by which topics are shared among the documents in the HDP and nHDP models. Each document in the HDP model is a different restaurant, and HDP is used for assigning words in document m to a local topic and for assigning local topics to one of the global topics. The discreteness of the random base distribution G_0 makes the M documents share some of their topics. In nHDP, a document chooses its way down the topic tree and is

distributed among all topics along its path down the tree. The HDPs in nHDP are used for assigning documents to sub-topics and for assigning sub-topics to topics. There is no sharing of topics among the restaurants. Topic sharing among documents is achieved through the structure of the topic tree. Documents share topics in the case of sharing some nodes on their paths.

APPROXIMATE INFERENCE

Given the training documents, we use Gibbs sampling (Geman and Geman, 1990), an example of the Markov chain Monte Carlo sampling method, to approximate the posterior distribution of hidden variables in the model. The section below addresses the conditional distributions of hidden variables in our generative model. The conditional distributions are used by the Gibbs sampler as proposal distributions.

We use the following distribution for sampling level assignment z_{mi} for word w_{mi}:

$$p(z_{mi} \mid \boldsymbol{z}^{-mi}, \boldsymbol{w}, \boldsymbol{j}, \boldsymbol{k}) \propto p(w_{mi} \mid \boldsymbol{w}^{-mi}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k}) p(z_{mi} \mid \boldsymbol{z}_m^{-i}), \tag{6}$$

where $\boldsymbol{z}$ is the set of level assignments for all words in the corpus, $\boldsymbol{w}$ is the set of all the words in the corpus, $\boldsymbol{j}=(\boldsymbol{j}_1, \boldsymbol{j}_2, \ldots, \boldsymbol{j}_M)$ is the set of path assignments for all documents in the corpus, $\boldsymbol{k}$ is the same as defined in Eq.(5), $p(w_{mi} \mid \boldsymbol{w}^{-mi}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k})$ is the likelihood of z_{mi} with respect to w_{mi}, and $p(z_{mi} \mid \boldsymbol{z}_m^{-i})$ is the conditional prior of z_{mi} determined by the Dirichlet multinomial distribution. When a superscript starting with a minus sign is attached to a variable, e.g., x^{-y}, the variable y is excluded from the set of variables represented by x; or y is omitted from the count number represented by x. In Eq.(6), $\boldsymbol{z}^{-mi}$ is the set of level assignments of all words in the corpus excluding z_{mi}, and $\boldsymbol{z}_m^{-i}$ is the set of level assignments of all words in document m excluding z_{mi}. By integrating out ε_m for document m and μ_k for each topic k,

$$p(z_{mi}=l \mid \boldsymbol{z}^{-mi}, \boldsymbol{w}, \boldsymbol{j}, \boldsymbol{k}) \propto \frac{n_l^{w_{mi}}+\eta}{n_l^*+W\eta} \frac{n_{ml}+\beta}{n_{m*}+L\beta}, \tag{7}$$

where $n_l^{w_{mi}}$ is the number of times that word w_{mi} has been assigned to the topic associated with the sub-topic on level l of path $\boldsymbol{j}_m$ excluding w_{mi}, n_l^* is the number of times that any word in the vocabulary has been assigned to that topic excluding w_{mi}, n_{ml} is the number of times that the words in document m have been assigned to level l excluding w_{mi}, and n_{m*} is the number of times that the words in document m have been assigned to all topics along path $\boldsymbol{j}_m$ excluding w_{mi}.

The conditional distribution used for sampling path assignment $\boldsymbol{j}_m$ of document m is

$$p(\boldsymbol{j}_m \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{j}^{-m}, \boldsymbol{k}) \propto p(\boldsymbol{w}_m \mid \boldsymbol{w}^{-m}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k}) p(\boldsymbol{j}_m \mid \boldsymbol{j}^{-m}). \tag{8}$$

The first term in the right-hand side of Eq.(8), as the likelihood of $\boldsymbol{j}_m$ with respect to $\boldsymbol{w}_m$, is calculated as follows:

$$p(\boldsymbol{w}_m \mid \boldsymbol{w}^{-m}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k}) = \prod_{l=1}^{L} \left(\frac{\Gamma(n_{j_m l}^* + W\eta)}{\prod_w \Gamma(n_{j_m l}^w + \eta)} \frac{\prod_w \Gamma(n_{j_m l}^w + n_{j_m lm}^w + \eta)}{\Gamma(n_{j_m l}^* + n_{j_m lm}^* + W\eta)} \right), \tag{9}$$

where $n_{j_m l}^*$ is the number of times that all words in the vocabulary have been assigned to the topic associated with the sub-topic on level l of path $\boldsymbol{j}_m$ excluding $\boldsymbol{w}_m$, $n_{j_m l}^{\boldsymbol{w}}$ is the number of times that word w has been assigned to that topic excluding $\boldsymbol{w}_m$, $n_{j_m lm}^*$ is the number of times that all words in document m have been assigned to that topic, and $n_{j_m lm}^{\boldsymbol{w}}$ is the number of times that word w in document m has been assigned to that topic.

The second term in the right-hand side of Eq.(8) is the prior of $\boldsymbol{j}_m$ as determined by Eqs.(4) and (5).

The conditional distribution used for sampling k_s for sub-topic s is

$$p(k_s \mid \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k}^{-s}) \propto p(\boldsymbol{w}_s \mid \boldsymbol{w}^{-s}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k}) p(k_s \mid \boldsymbol{k}^{-s}), \tag{10}$$

where $\boldsymbol{w}_s$ is the set of all words assigned to sub-topic s, $p(\boldsymbol{w}_s|\boldsymbol{w}^{-s}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{k})$ is the likelihood of k_s with respect to $\boldsymbol{w}_s$, and $p(k_s|\boldsymbol{k}^{-s})$ is determined by Eq.(5).

EXPERIMENT AND DISCUSSION

In this section, we demonstrate the effectiveness of the nHDP model with both synthetic and the 20-newsgroup dataset, which is widely used in topic modeling research for demonstrating the effectiveness of algorithms (Bast and Majumdar, 2005; Mimno *et al*., 2007), and compare the results with those of the hLDA model.

A corpus of four hundred and fifty 100-word documents was generated using the model in Fig.2. Each node in the tree was a 5×5 image. We treated each of the 25 pixels as a word in a vocabulary and each node in the tree as a topic. For each document, we sampled its path from one of the four paths uniformly, and a topic mixing proportion from a Dirichlet distribution. For each word in a document, we sampled one topic from the mixing proportion and extracted the word from that topic.

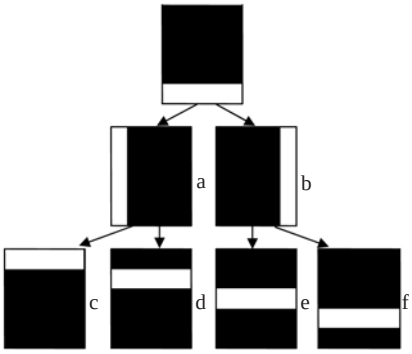

Fig.2 Model used in the bar example

We implemented the hLDA model with Gibbs sampling in the Java programming language. We ran hLDA on the data for 5000 iterations, took one sample, and restarted the chain several times. The log likelihood $p(\boldsymbol{w}|\boldsymbol{c},\boldsymbol{z})$ in hLDA usually stabilizes in the first 100 iterations. Fig.3 is one result obtained after 5000 iterations. It is an interesting demonstration of two local maxima problems in hLDA that happens commonly in our experiments.

Duplicated parent topic problem

The nodes a and c in Fig.3 should have been merged into one node, so should the nodes e and g. Path assignment of each document in hLDA can only be sampled separately when implemented with Gibbs sampling. But path assignments for documents on paths ab and cd are strongly related. The only way to merge nodes a and c in Fig.3 is to move the documents assigned to path cd to a path ad′ (d′ is a new child node of node a) one by one, or vice versa. However, when sampling the path assignment for a document assigned to path cd in the previous iteration, the conditional probability of path ad′ can be much lower than that of path cd. That is because the other similar documents on path cd make the likelihood of the document being assigned to path cd much higher. Unless the nested CRP gives path ad′ a prior probability a lot higher than that of path cd (Note: this happens when there are much more documents assigned to node a than to node c), a document being moved to path ad′ from path cd will return to path cd after one iteration or two. One way for hLDA to avoid being trapped in local maxima like this is to form the correct tree structure during early iterations of the Markov chain when there are only a few documents on each path and small differences in likelihood between paths.

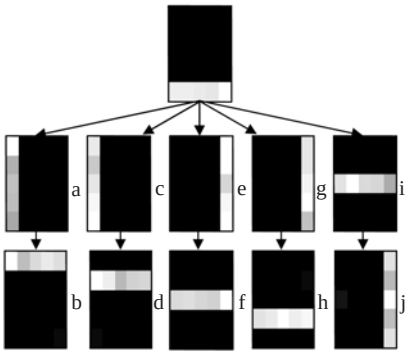

Fig.3 One result of hLDA

Inverted path problem

Path ij in Fig.3 is path ef with its level-2 and level-3 topics inverted. This happens for two reasons: (1) The hidden random variables are randomly initialized. The documents from the same path in the true model can have their words initialized with different level assignments. (2) The conditional distribution used in the Gibbs sampler fixes level assignments $\boldsymbol{z}_m$ of the words in a document m when sampling path assignment c_m for the document. So path ef will have a quite low likelihood compared to path ij when sampling path assignment for a document assigned to path ij in the previous iteration.

The nHDP model was designed with the first problem in mind. We replace the CRP in the hLDA model with an HDP. The two-layered clustering (Note: a sub-topic is a cluster of documents, and a topic is a

cluster of sub-topics) makes the nHDP a good model to deal with the duplicated parent topic problem. Sampling a topic for a sub-topic according to Eq.(10) is equivalent to sampling a path for a group of documents together. Thus, when transferring a document from one path to another, we move all other similar documents assigned to the same path with it. This solves the duplicated parent topic problem.

The nHDP model can still be trapped in local maxima as described in the inverted path problem. To deal with this problem, we introduce the mapped nHDP (mnHDP) model, an extension to the nHDP model. The generative model is described as follows:

For each document m:

Step 1: Let j_{m1} be the root topic.

Step 2: For each level $l \in \{2, 3, \ldots, L\}$: (i) Draw a sub-topic according to Eq.(4), and set j_{ml} to be the sub-topic; (ii) If a new sub-topic is created in step (i), draw a topic according to Eq.(5), attain parameters of the multinomial distribution μ for the new topic from $Dir(\eta)$, and assign the sub-topic to it.

Step 3: Draw an L-dimensional topic mixing proportion vector $\boldsymbol{\varepsilon}_m$ from $Dir(\beta)$.

Step 4: Draw level mapping $\boldsymbol{m}_m$ from a prior distribution.

Step 5: For each word w_{mi} in document m: (i) Draw level assignment $z_{mi} \in \{1, 2, \ldots, L\}$ from $Mult(\boldsymbol{\varepsilon}_m)$. (ii) Map z_{mi} to the mapped assignment z'_{mi} according to the level mapping $\boldsymbol{m}_m$. (iii) If $z'_{mi}=1$, draw w_{mi} from the root topic; otherwise, from the topic to which the sub-topic on level z'_{mi} of path $\boldsymbol{j}_m$ is assigned.

Instead of assigning w_{mi} to a level according to the level assignment z_{mi}, we mapped z_{mi} to z'_{mi} according to a document-specific level mapping $\boldsymbol{m}_m$. For example, in a three-level topic tree, $\boldsymbol{m}_m=(2, 3, 1)$ means that, a word with level assignment $z_{mi}=1$ is actually assigned to level 2 on path $\boldsymbol{j}_m$, a word with $z_{mi}=2$ to level 3, and a word with $z_{mi}=3$ to level 1. A five-word document with $\boldsymbol{z}_m=(1, 3, 2, 1, 2, 3)$ and $\boldsymbol{m}_m=(2, 3, 1)$ will have its words assigned to levels (2, 1, 3, 2, 3, 1). We added the constraint that $\boldsymbol{m}_m$ is a one-to-one mapping. The sample space of $\boldsymbol{m}_m$ is usually small: on a three-level topic tree, the sample space size is only 2 if we further add the constraint that words with $z_{mi}=1$ are always mapped to level 1. With such a small sample space, we can jointly sample $\boldsymbol{m}_m$ with path assignment $\boldsymbol{j}_m$ as in Eq.(11) to help the Markov chain escape local maxima:

$$p(\boldsymbol{j}_m, \boldsymbol{m}_m \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{j}^{-m}, \boldsymbol{m}^{-m}, \boldsymbol{k}) \propto p(\boldsymbol{w}_m \mid \boldsymbol{w}^{-m}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{m}, \boldsymbol{k}) p(\boldsymbol{j}_m \mid \boldsymbol{j}^{-m}) p(\boldsymbol{m}_m), \tag{11}$$

where $p(\boldsymbol{w}_m|\boldsymbol{w}^{-m}, \boldsymbol{z}, \boldsymbol{j}, \boldsymbol{m}, \boldsymbol{k})$ is the likelihood of $(\boldsymbol{j}_m, \boldsymbol{m}_m)$ with respect to $\boldsymbol{w}_m$. It is similar to Eq.(9) except that we add the level mapping. $p(\boldsymbol{m}_m)$ is the prior of the document-specific level mapping. We set the probability of a document with its level-2 and level-3 inverted to 1/30 in our experiments of modeling topic trees of three levels.

In another bar example, we constructed a three-level hierarchical topic tree with 3 nodes on level 2, with 5, 2, 2 child nodes on level 3 respectively. Eight hundred 200-word documents were generated similarly as in the previous bar example. We ran 4 topic models—hLDA, mhLDA (the hLDA model enhanced with level mapping), nHDP, and mnHDP—on the dataset. For each topic model, we let 50 independent randomly initialized Markov chains burn in for 2000 iterations and took one sample from each chain. To test how different $\boldsymbol{\eta}$ values affect the effectiveness of the 4 topic models, especially the ability to correctly merge level-2 topics, we conducted the above experiment twice, one with $\boldsymbol{\eta}=(0.1, 0.1, 0.1)$, the other with $\boldsymbol{\eta}=(0.1, 1.0, 0.1)$ [$\boldsymbol{\eta}=(0.1, 1.0, 0.1)$ means that parameters of all level-1 and level-3 topics are sampled from a symmetric Dirichlet distribution with $\eta=0.1$ and parameters of all level-2 topics are sampled from a symmetric Dirichlet distribution with $\eta=1.0$]. Table 1 shows the number of errors averaged over 50 samples for each topic model with two different $\boldsymbol{\eta}$ values. We can see that a larger $\boldsymbol{\eta}$ value can help the hLDA model merge topics at level 2. The nHDP model and the mnHDP model are more capable of dealing with the duplicated parent problem and the

Table 1 Average error numbers of the four topic models

Model	Average number of errors					
	$\boldsymbol{\eta}=(0.1, 0.1, 0.1)$			$\boldsymbol{\eta}=(0.1, 1.0, 0.1)$		
	DP	IP	O	DP	IP	O
hLDA	3.70	0.80	2.40	1.52	0.02	1.28
mhLDA	3.80	0.00	2.42	1.42	0.00	1.34
nHDP	0.10	0.24	2.52	0.00	0.02	1.06
mnHDP	0.08	0.00	2.54	0.00	0.00	1.38

Each value in the table is the average number of errors over 50 samples for a topic model. DP: number of duplicated parent errors; IP: number of inverted path errors; O: all other errors. The true data model is a 3-level hierarchical topic tree with 3 nodes on level 2, with 5, 2, 2 child nodes on level 3, respectively

inverted path problem with both $\boldsymbol{\eta}$ values. Level mapping can help both hLDA and nHDP escape local maxima as described in the inverted path problem.

To test our model in real world problems, we generated a sample of 1105 documents from the well-known 20-newsgroup dataset by uniformly sampling documents from each newsgroup. After removing stop words from the vocabulary, we acquired a total of 93975 words with a vocabulary size of 2702. We ran 4 topic models—hLDA, mhLDA, nHDP, and mnHDP—on the dataset. After burning the Markov chain for 10000 iterations, we took 3 samples with a gap of 1000 iterations. We restarted all chains 5 times. In all models, we set η=0.5 for the symmetric Dirichlet distribution for all topics (when η=0.1 all four topic models produced over 300 topics, and the topics are hard to interpret), while the CRP parameter in Eq.(1) for hLDA and mhLDA, α in Eq.(4), and γ in Eq.(5) were all set to 1. The Dirichlet prior for word level assignments was set to (10, 6, 3) in both models. It took 8 h to train the mnHDP model on a workstation with a 2.33 GHz Intel CPU and 2 GB memory. To determine if the model has stabilized, we calculated the log likelihood of the 4 models with respect to the training data by using the empirical likelihood (Li and McCallum, 2006). The log likelihood of all the 4 models stabilized in the first 100 iterations.

mnHDP produced a much more compact tree structure than hLDA did. Fig.4 is a histogram for the size of level-2 topics (number of documents assigned to level-2 topics) in hLDA and mnHDP. The *x*-axis is the number of documents assigned to a level-2 topic, and the *y*-axis is the number of level-2 topics of a certain size. There are 42 level-2 topics in the hLDA topic tree, as compared with only 16 in mnHDP. The average number of documents contained in a level-2 topic in hLDA is 26, whereas it is 69 in nHDP.

Many newsgroups in the 20-newsgroup dataset are similar in word usage. We removed stop words, i.e., a standard stop word list and the top 200 words in the root topic in one of the samples, from the hLDA model (the root topic contains stop words specific to the newsgroups, such as subject, lines, writes, article, quote), and kept only the top 400 words with highest average mutual information with the original newsgroup label. For each newsgroup, we merged all the contained articles into one document, and represented the newsgroup with that document in the vector space model. Fig.5 shows the inter-newsgroup similarity measured by Jensen-Shannon divergence (Lin, 1991). According to Fig.5, we divided the 20 newsgroups into 8 partitions: (comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, misc.forsale, comp.windows.x, comp.sys.mac.hardware, sci.electronics), (talk.politics.guns, talk.politics.mideast, talk.politics.misc), (talk.religion.misc, soc.religion.christian, alt.atheism), (rec.autos, rec.motorcycles), (rec.sport.baseball, rec.sport.hockey); the other 3 newsgroups, sci.med, sci.space, and sci.crypt, showing little inter-newsgroup similarity, were put in 3 separate partitions.

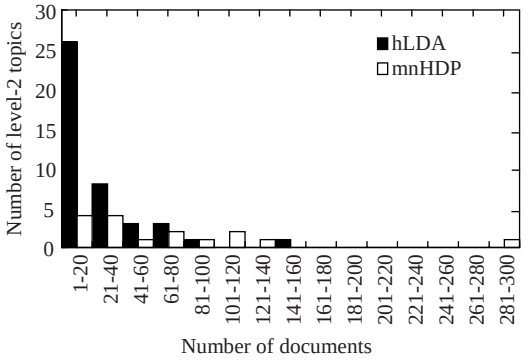

Fig.4 Histogram for documents number on level-2 topics
The *x*-axis is the number of documents assigned to a level-2 topic and the *y*-axis is the number of level-2 topics of a certain size. mnHDP produced a much more compact tree structure

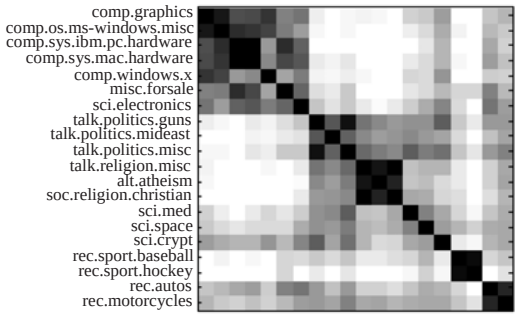

Fig.5 Jensen-Shannon (JS) divergence of the 20-newsgroup dataset
The *y*-axis lists the newsgroup names, and newsgroup names on the *x*-axis are in the same order (the leftmost is comp.graphics and the rightmost is rec.motorcycles). We can see one large partition about the computer, and 7 other smaller partitions

Both nHDP and hLDA can capture the large partition about the computer (Fig.5) with a level-2 topic. But hLDA is not good at capturing

inter-newsgroup similarity shared by only a small number of documents in its level-2 topics like topics 1 and 2 in Fig.6a. Such small partitions usually have separate level-2 topics in topic trees produced by hLDA in our experiments (Fig.6b).

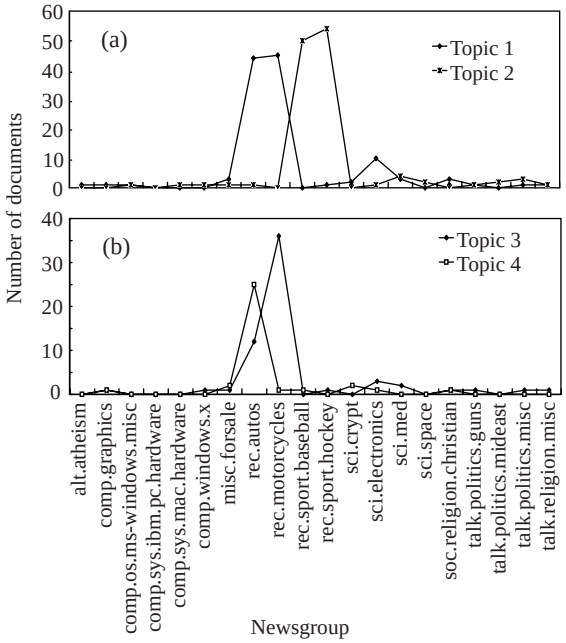

Fig.6 Distribution over newsgroups for two level-2 topics discovered by mnHDP (a) and hLDA (b)
Topics 1 and 2 in (a) correspond to the partitions (rec.autos, rec.motorcycles) and (rec.sport.baseball, rec.sport.hockey) respectively, and topics 3 and 4 in (b) correspond to the split (rec.autos, rec.motorcycles) partition in Fig.5

Because mutual information does not penalize results with many similar topics (Manning *et al*., 2008), we measured how well the level-2 topics produced by different topic models correspond to the 8 partitions shown in Fig.5 by normalized mutual information (NMI) (Press *et al*., 1992):

$$NMI(t, p) = \frac{MI(t, p)}{H(t)}, \tag{12}$$

where t is the random variable that represents the level-2 topics produced by the topic models, p is the random variable that indicates the partitions in Fig.5, $MI(t, p)$ is the mutual information between t and p, and $H(t)$ is the entropy of t. The normalized mutual information between two random variables is between 0 and 1. As shown in Table 2, the nHDP model is 30% better than hLDA with respect to normalized mutual information, and level mapping can further improve both models.

Following (Li and McCallum, 2006), we checked the predictive power of our model with hLDA by calculating the empirical likelihood (EL) of 378 held-out documents from the 20-newsgroup dataset. nHDP and mnHDP improved the empirical likelihood by 1% (Table 2).

Table 2 *NMI* and *EL* of the four topic models

Model	*NMI*	*EL*
hLDA	0.348 625	−236 440
mhLDA	0.392 895	−236 321
nHDP	0.453 422	−234 364
mnHDP	0.597 039	−234 177

NMI: normalized mutual information as defined in Eq.(12); *EL*: empirical likelihood of 378 held-out documents

Fig.7 shows part of the topic tree produced by mnHDP. The leftmost level-2 topic corresponds to the large partition related to the computer in Fig.5, while the rightmost level-2 topic corresponds to the partition (rec.autos, rec.motorcycles).

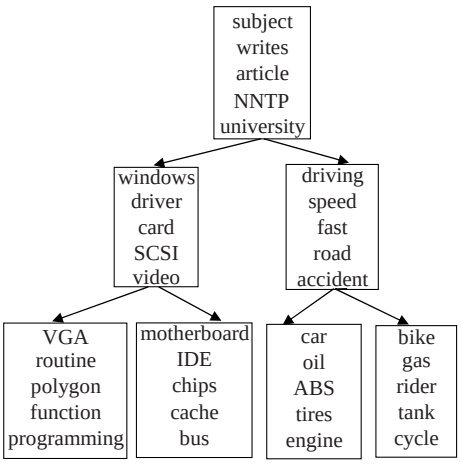

Fig.7 Words with high probabilities in part of the topic tree produced by mnHDP

CONCLUSION

We extended the hierarchical Dirichlet process to model latent hierarchical topic structures embedded in text corpora. With a specially designed two-layered clustering, our model is more likely to

escape local maxima than the hLDA model. We implemented the model with Gibbs sampling in the Java programming language. Experiments on both synthetic and real dataset show that, our model not only produces a more compact tree structure, but also is able to capture fine-grained topic relationships which are often missed by hLDA.

There are several possible extensions to the model. First, we can extend the model by allowing tree paths of the topic tree to have different lengths. This extension is useful when part of the corpus has a richer hierarchical topic structure than the rest. Second, we would like to see if a better sampling algorithm can deal with the duplicated parent problem and the inverted path problem. Third, we can enrich the model by introducing more prior knowledge, such as threading relationship among the documents in newsgroups.

References

Bast, H., Majumdar, D., 2005. Why Spectral Retrieval Works. Proc. 28th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.11-18. [doi:10.1145/1076034.1076040]

Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Polya Urn schemes. *Ann. Statist.*, **1**(2):353-355. [doi:10.1214/aos/1176342372]

Blei, D.M., Lafferty, J.D., 2006. Dynamic Topic Models. Proc. 23rd Int. Conf. on Machine Learning, p.113-120. [doi:10.1145/1143844.1143859]

Blei, D.M., Lafferty, J.D., 2007. A correlated topic model of science. *Ann. Appl. Statist.*, **1**(1):17-35. [doi:10.1214/07-AOAS114]

Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B., 2003a. Hierarchical Topic Models and the Nested Chinese Restaurant Process. NIPS, p.17-24.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003b. Latent Dirichlet allocation. *J. Mach. Learning Res.*, **3**(4-5):993-1022. [doi:10.1162/jmlr.2003.3.4-5.993]

Boley, D.L., 1998. Principal direction divisive partitioning. *Data Min. Knowl. Discov.*, **2**(4):325-344. [doi:10.1023/A:1009740529316]

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, **41**(6):391-407. [doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9]

Dhillon, I.S., Modha, D.S., 2001. Concept decompositions for large sparse text data using clustering. *Mach. Learning*, **42**(1/2):143-175. [doi:10.1023/A:1007612920971]

Elkan, C., 2006. Clustering Documents with an Exponential-family Approximation of the Dirichlet Compound Multinomial Distribution. Proc. 23rd Int. Conf. on Machine Learning, p.289-296. [doi:10.1145/1143844.1143881]

Geman, S., Geman, D., 1990. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *In*: Shafer, G., Pearl, J. (Eds.), Readings in Uncertain Reasoning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p.452-472.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. *PNAS*, **101**(Suppl. 1):5228-5235. [doi:10.1073/pnas.0307752101]

Li, W., McCallum, A., 2006. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations. Proc. 23rd Int. Conf. on Machine Learning, p.577-584. [doi:10.1145/1143844.1143917]

Lin, J., 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, **37**(1):145-151. [doi:10.1109/18.61115]

Manning, C.D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.

Mimno, D., Li, W., McCallum, A., 2007. Mixtures of Hierarchical Topics with Pachinko Allocation. Proc. 24th Int. Conf. on Machine Learning, p.633-640. [doi:10.1145/1273496.1273576]

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1992. Numerical Recipes in C. Cambridge University Press, Cambridge, England.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The Author-topic Model for Authors and Documents. Proc. 20th Conf. on Uncertainty in Artificial Intelligence, p.487-494.

Strehl, A., Ghosh, J., Mooney, R., 2000. Impact of Similarity Measures on Web-page Clustering. Proc. 17th National Conf. on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search, p.58-64.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. *J. Am. Statist. Assoc.*, **101**(476):1566-1581. [doi:10.1198/016214506000000302]

Walker, D.D., Ringger, E.K., 2008. Model-based Document Clustering with a Collapsed Gibbs Sampler. Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.704-712. [doi:10.1145/1401890.1401975]

Wallach, H.M., 2006. Topic Modeling: Beyond Bag-of-words. Proc. 23rd Int. Conf. on Machine Learning, p.977-984. [doi:10.1145/1143844.1143967]

Wei, X., Croft, B.W., 2006. LDA-based Document Models for Ad-hoc Retrieval. Proc. 29th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.178-185. [doi:10.1145/1148170.1148204]

Zhang, Z., Phan, X.H., Horiguchi, S., 2008. An Efficient Feature Selection Using Hidden Topic in Text Categorization. Proc. 22nd Int. Conf. on Advanced Information Networking and Applications, p.1223-1228. [doi:10.1109/WAINA.2008.137]