SSHLDA: A Semi-Supervised Hierarchical Topic Model

Xian-Ling Mao[♠]*, Zhao-Yan Ming[♡], Tat-Seng Chua[♡], Si Li[♣], Hongfei Yan[♠]†, Xiaoming Li[♠]
[♠]Department of Computer Science and Technology, Peking University, China
[♡]School of Computing, National University of Singapore, Singapore
[♣]School of ICE, Beijing University of Posts and Telecommunications, China
{xianlingmao,lxm}@pku.edu.cn, yhf@net.pku.edu.cn
{chuats,mingzhaoyan}@nus.edu.sg, lisi@bupt.edu.cn

Abstract

Supervised hierarchical topic modeling and unsupervised hierarchical topic modeling are usually used to obtain hierarchical topics, such as hLLDA and hLDA. Supervised hierarchical topic modeling makes heavy use of the information from observed hierarchical labels, but cannot explore new topics; while unsupervised hierarchical topic modeling is able to detect automatically new topics in the data space, but does not make use of any information from hierarchical labels. In this paper, we propose a semi-supervised hierarchical topic model which aims to explore new topics automatically in the data space while incorporating the information from observed hierarchical labels into the modeling process, called *Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA)*. We also prove that hLDA and hLLDA are special cases of SSHLDA. We conduct experiments on Yahoo! Answers and ODP datasets, and assess the performance in terms of perplexity and clustering. The experimental results show that predictive ability of SSHLDA is better than that of baselines, and SSHLDA can also achieve significant improvement over baselines for clustering on the FScore measure.

1 Introduction

Topic models, such as latent Dirichlet allocation (LDA), are useful NLP tools for the statistical analysis of document collections and other discrete data. Furthermore, hierarchical topic modeling is able to obtain the relations between topics — parent-child and sibling relations. Unsupervised hierarchical topic modeling is able to detect automatically new topics in the data space, such as hierarchical Latent Dirichlet Allocation (hLDA) (Blei et al., 2004). hLDA makes use of nested Dirichlet Process to automatically obtain a L-level hierarchy of topics. Modern Web documents, however, are not merely collections of words. They are usually documents with hierarchical labels – such as Web pages and their placement in hierarchical directories (Ming et al., 2010). Unsupervised hierarchical topic modeling cannot make use of any information from hierarchical labels, thus supervised hierarchical topic models, such as *hierarchical Labeled Latent Dirichlet Allocation* (hLLDA) (Petinot et al., 2011), are proposed to tackle this problem. hLLDA uses hierarchical labels to automatically build corresponding topic for each label, but it cannot find new latent topics in the data space, only depending on hierarchy of labels.

As we know that only about 10% of an iceberg's mass is seen outside while about 90% of it is unseen, deep down in water. We think that a corpus with hierarchical labels should include not only observed topics of labels, but also there are more latent topics, just like icebergs. hLLDA can make use of the information from labels; while hLDA can explore latent topics. How can we combine the merits of the two types of models into one model?

An intuitive and simple combinational method is like this: first, we use hierarchy of labels as basic hierarchy, called Base Tree (BT); then we use hLDA to build automatically topic hierarchy for each leaf

*This work was done in National University of Singapore.
†Corresponding author.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 800–809, Jeju Island, Korea, 12–14 July 2012.

node in BT, called Leaf Topic Hierarchy (LTH); finally, we add each LTH to corresponding leaf in the BT and obtain a hierarchy for the entire dataset. We refer the method as Simp-hLDA. The performance of the Simp-hLDA is not so good, as can be seen from the example in Figure 3 (b). The drawbacks are: (i) the leaves in BT do not obtain reasonable and right words distribution, such as "Computers & Internet" node in Figure 3 (b), its topical words, "the to you and a", is not about "Computers & Internet"; (ii) the non-leaf nodes in BT cannot obtain words distribution, such as "Health" node in Figure 3 (b); (iii) it is a heuristic method, and thus Simp-hLDA has no solid theoretical basis.

To tackle the above drawbacks, we explore the use of probabilistic models for such a task where the hierarchical labels are merely viewed as a part of a hierarchy of topics, and the topics of a path in the whole hierarchy generate a corresponding document. Our proposed generative model learns both the latent topics of the underlying data and the labeling strategies in a joint model, by leveraging on the hierarchical structure of labels and Hierarchical Dirichlet Process.

We demonstrate the effectiveness of the proposed model on large, real-world datasets in the question answering and website category domains on two tasks: the topic modeling of documents, and the use of the generated topics for document clustering. Our results show that our joint, semi-hierarchical model outperforms the state-of-the-art supervised and unsupervised hierarchical algorithms. The contributions of this paper are threefold: (1) We propose a joint, generative semi-supervised hierarchical topic model, i.e. Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA), to overcome the defects of hLDA and hLLDA while combining the their merits. SSHLDA is able to not only explore new latent topics in the data space, but also makes use of the information from the hierarchy of observed labels; (2) We prove that hLDA and hLLDA are special cases of SSHLDA; (3) We develop a gibbs sampling inference algorithm for the proposed model.

The remainder of this paper is organized as follows. We review related work in Section 2. In Section 3, we introduce some preliminaries; while we introduce *SSHLDA* in Section 4. Section 5 details a gibbs sampling inference algorithm for SSHLDA; while Section 6 presents the experimental results. Finally, we conclude the paper and suggest directions for future research in Section 7.

2 Related Work

There have been many variations of topic models. The existing topic models can be divided into four categories: *Unsupervised non-hierarchical topic models, Unsupervised hierarchical topic models*, and their corresponding supervised counterparts.

Unsupervised non-hierarchical topic models are widely studied, such as LSA (Deerwester et al., 1990), pLSA (Hofmann, 1999), LDA (Blei et al., 2003), Hierarchical-concept TM (Chemudugunta et al., 2008c; Chemudugunta et al., 2008b), Correlated TM (Blei and Lafferty, 2006) and Concept TM (Chemudugunta et al., 2008a; Chemudugunta et al., 2008b) etc. The most famous one is Latent Dirichlet Allocation (LDA). LDA is similar to pLSA, except that in LDA the topic distribution is assumed to have a Dirichlet prior. LDA is a completely unsupervised algorithm that models each document as a mixture of topics. Another famous model that not only represents topic correlations, but also learns them, is the Correlated Topic Model (CTM). Topics in CTM are not independent; however it is noted that only pairwise correlations are modeled, and the number of parameters in the covariance matrix grows as the square of the number of topics.

However, the above models cannot capture the relation between super and sub topics. To address this problem, many models have been proposed to model the relations, such as Hierarchical LDA (HLDA) (Blei et al., 2004), Hierarchical Dirichlet processes (HDP) (Teh et al., 2006), Pachinko Allocation Model (PAM) (Li and McCallum, 2006) and Hierarchical PAM (HPAM) (Mimno et al., 2007) etc. The relations are usually in the form of a hierarchy, such as the tree or Directed Acyclic Graph (DAG). Blei et al. (2004) proposed the hLDA model that simultaneously learns the structure of a topic hierarchy and the topics that are contained within that hierarchy. This algorithm can be used to extract topic hierarchies from large document collections.

Although unsupervised topic models are suffi-

ciently expressive to model multiple topics per document, they are inappropriate for labeled corpora because they are unable to incorporate the observed labels into their learning procedure. Several modifications of LDA to incorporate supervision have been proposed in the literature. Two such models, Supervised LDA (Blei and McAuliffe, 2007; Blei and McAuliffe, 2010) and DiscLDA (Lacoste-Julien et al., 2008) are first proposed to model documents associated only with a single label. Another category of models, such as the MM-LDA (Ramage et al., 2009b), Author TM (Rosen-Zvi et al., 2004), Flat-LDA (Rubin et al., 2011), Prior-LDA (Rubin et al., 2011), Dependency-LDA (Rubin et al., 2011) and Partially LDA (PLDA) (Ramage et al., 2011) etc., are not constrained to one label per document because they model each document as a bag of words with a bag of labels. However, these models obtain topics that do not correspond directly with the labels. Labeled LDA (LLDA) (Ramage et al., 2009a) can be used to solve this problem.

None of these non-hierarchical supervised models, however, leverage on dependency structure, such as parent-child relation, in the label space. For hierarchical labeled data, there are also few models that are able to handle the label relations in data. To the best of our knowledge, only hLLDA (Petinot et al., 2011) and HSLDA (Perotte et al., 2011) are proposed for this kind of data. HSLDA cannot obtain a probability distribution for a label. Although hLLDA can obtain a distribution over words for each label, hLLDA is unable to capture the relations between parent and child node using parameters, and it also cannot detect automatically latent topics in the data space. In this paper, we will propose a generative topic model to tackle these problems of hLLDA.

3 Preliminaries

The nested Chinese restaurant process (nCRP) is a distribution over hierarchical partitions (Blei et al., 2004). It generalizes the Chinese restaurant process (CRP), which is a distribution over partitions. The CRP can be described by the following metaphor. Imagine a restaurant with an infinite number of tables, and imagine customers entering the restaurant in sequence. The d^{th} customer sits at a table according to the following distribution,

Table 1: Notations used in the paper.

Sym	Description
V	Vocabulary (word set), w is a word in V
D	Document collection
T_j	The set of paths in the sub-tree whose root is the j^{th} leaf node in the hierarchy of observed topics
m	A document m that consists of words and labels
$\mathbf{w}_m$	The text of document m, w_i is i^{th} words in $\mathbf{w}$
$\mathbf{c}_m$	The topic set of document m
$\mathbf{c}_{o_m}$	The set of topics with observed labels for document m
$\mathbf{c}_{e_m}$	The set of topics without labels for document m
$\boldsymbol{c}_{e_{-m}}$	The set of latent topics for all documents other than m
$\boldsymbol{z}_{e_m}$	The assignment of the words in the m^{th} document to one of the latent topics
$\boldsymbol{w}_{e_m}$	The set of the words belonging to one of the latent topics in the the m^{th} document
$z_{m,n}$	The assignment of the n^{th} word in the m^{th} document to one of the L available topics
$\mathbf{z}$	The set of $z_{m,n}$ for all words in all documents
c_i	A topic in the i^{th} level in the hierarchy
θ	The word distribution set for Z, i.e., $\{\theta\}_{z\in\mathbf{c}}$
α	Dirichlet prior of θ
δ_{c_i}	The multinomial distribution over the sub-topics of c_{i-1}
μ_{c_i}	Dirichlet prior of δ_{c_i}
η	Dirichlet prior of β
β	The multinomial distribution of words
θ_m	The distributions over topics for document m
$\boldsymbol{\theta}$	The set for θ_m, $m \in \{1, ..., D\}$

$$p(c_d = k|c_{1:(d-1)}) \propto \begin{cases} m_k & \text{if } k \text{ is previous occupied} \\ \gamma & \text{if } k \text{ is a new tabel,} \end{cases} \quad (1)$$

where m_k is the number of previous customers sitting at table k and γ is a positive scalar. After D customers have sat down, their seating plan describes a partition of D items.

In the nested CRP, imagine now that tables are organized in a hierarchy: there is one table at the first level; it is associated with an infinite number of tables at the second level; each second-level table is associated with an infinite number of tables at the third level; and so on until the L^{th} level. Each customer enters at the first level and comes out at the L^{th} level, generating a path with L tables as she sits in each restaurant. Moving from a table at level l to one of its subtables at level $l+1$, the customer draws following the CRP using Formula (1). In this paper, we will make use of nested CRP to explore latent topics in data space.

To elaborate our model, we first define two concepts. If a model can learn a distribution over words for a label, we refer the topic with a corresponding label as a **labeled topic**. If a model can learn an unseen and latent topic without a label, we refer the

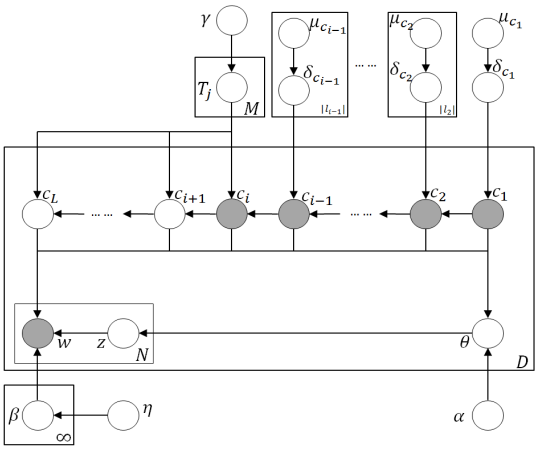

Figure 1: The graphical model of SSHLDA.

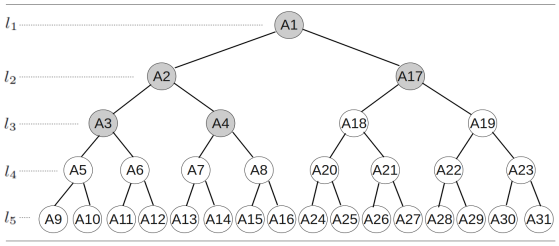

Figure 2: One illustration of SSHLDA. The tree has 5 levels. The shaded nodes are observed topics, and circled nodes are latent topics. The latent topics are generated automatically by SSHLDA model. After learning, each node in this tree will obtain a corresponding probability distribution over words, i.e. a topic.

topic as a **latent topic**.

4 The Semi-Supervised Hierarchical Topic Model

In this section, we will introduce a semi-supervised hierarchical topic model, i.e., the *Semi-Supervised Hierarchical Latent Dirichlet Allocation* (SSHLDA). SSHLDA is a probabilistic graphical model that describes a process for generating a hierarchical labeled document collection. Like hierarchical Labeled LDA (hLLDA) (Petinot et al., 2011), SSHLDA can incorporate labeled topics into the generative process of documents. On the other hand, like hierarchical Latent Dirichlet Allocation (hLDA) (Blei et al., 2004), SSHLDA can automatically explore latent topic in data space, and extend the existing hierarchy of observed topics. SSHLDA makes use of not only observed topics, but also latent topics.

The graphical model of SSHLDA is illustrated in Figure 1. In the model, N is the number of words in a document, D is the total number of documents in a collection, M is the number of leaf nodes in hierarchical observed nodes, c_i is a node in the i^{th} level in the hierarchical tree, η, α and μ_{c_i} are dirichlet prior parameters, β_k is a distribution over words, θ is a document-specific distribution over topics, δ_{c_i} is a multinomial distribution over observed sub-topics of topic c_i, w is an observed word, z is the topic assigned to w, $Dir_k(.)$ is a k-dimensional Dirichlet distribution, T_j is a set of paths in the hierarchy of latent topics for j^{th} leaf node in the hierarchy of observed topics, γ is a Multi-nomial distribution over paths in the tree. All notations used in this paper are listed in Table 1.

SSHLDA, as shown in Figure 1, assumes the following generative process:

(1) For each table $k \in T$ in the infinite tree,
  (a) Draw a topic $\beta_k \sim Dir(\eta)$.

(2) For each document, $m \in \{1, 2, ..., D\}$
  (a) Let c_1 be the root node.
  (b) For each level $l \in \{2, ..., L\}$:
    (i) If nodes in this level have been observed, draw a node c_l from $Mult(\delta_{c_{l-1}}|\mu_{c_{l-1}})$.
    (ii) Otherwise, draw a table c_l from restaurant c_{l-1} using Formula (1).
  (c) Draw an L-dimensional topic proportion vector θ_m from $Dir(\alpha)$.
  (d) For each word $n \in \{1, ..., N\}$:
    (i) Draw $z \in \{1, ..., L\}$ from $Mult(\theta)$.
    (ii) Draw w_n from the topic associated with restaurant c_z.

As the example showed in Figure 2, we assume that we have known a hierarchy of observed topics: {A1,A2,A17,A3,A4}, and assume the height of the desired topical tree is $L = 5$. All circled nodes are latent topics, and shaded nodes are observed topics. A possible generative process for a document *m* can be: It starts from A1, and chooses node A17 at level 2, and then chooses A18, A20 and A25 in the following levels. Thus we obtain a path: $\boldsymbol{c}_m = \{A1, A17, A18, A20, A25\}$. After getting the path for m, SSHLDA generates each word from one of topics in this set of topics $\boldsymbol{c}_m$.

5 Probabilistic Inference

In this section, we describe a Gibbs sampling algorithm for sampling from the posterior and corresponding topics in the SSHLDA model. The Gibbs sampler provides a method for simultaneously exploring the model parameter space (the latent topics of the whole corpus) and the model structure space (L-level trees).

In SSHLDA, we sample the paths $\boldsymbol{c}_m$ for document m and the per-word level allocations to topics in those paths $z_{m,n}$. Thus, we approximate the posterior $p(\boldsymbol{c}_m, \boldsymbol{z}_m|\gamma, \eta, \boldsymbol{w}, \boldsymbol{\mu})$. The hyper-parameter γ reflects the tendency of the customers in each restaurant to share tables, η denotes the expected variance of the underlying topics (e.g., $\eta \ll 1$ will tend to choose topics with fewer high-probability words), μ_{c_i} is the dirichlet prior of δ_{c_i}, and $\boldsymbol{\mu}$ is the set of μ_{c_i}. $w_{m,n}$ denotes the n^{th} word in the m^{th} document; and $c_{m,l}$ represents the restaurant corresponding to the l^{th}-level topic in document m; and $z_{m,n}$, the assignment of the n^{th} word in the m^{th} document to one of the L available topics. All other variables in the model, θ and β, are integrated out. The Gibbs sampler thus assesses the values of $z_{m,n}$ and $c_{m,l}$.

The Gibbs sampler can be divided into two main steps: the sampling of level allocations and the sampling of path assignments.

First, given the values of the SSHLDA hidden variables, we sample the $c_{m,l}$ variables which are associated with the CRP prior. Noting that $\boldsymbol{c}_m$ is composed of $\boldsymbol{c}_{o_m}$ and $\boldsymbol{c}_{e_m}$, $\boldsymbol{c}_{o_m}$ is the set of observed topics for document m, and $\boldsymbol{c}_{e_m}$ is the set of latent topics for document m. The conditional distribution for $\boldsymbol{c}_m$, the L topics associated with document m, is:

$$\begin{aligned} &p(\boldsymbol{c}_m|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{c}_{-m}, \boldsymbol{\mu}) \\ &\quad = p(\boldsymbol{c}_{o_m}|\boldsymbol{\mu}) p(\boldsymbol{c}_{e_m}|\boldsymbol{z}_{e_m}, \boldsymbol{w}_{e_m}, \boldsymbol{c}_{e_{-m}}) \\ &\quad \propto p(\boldsymbol{c}_{o_m}|\boldsymbol{\mu}) p(\boldsymbol{w}_{e_m}|\boldsymbol{c}_{e_m}, \boldsymbol{w}_{e_{-m}}, \boldsymbol{z}_{e_m}) \\ &\quad\quad p(\boldsymbol{c}_{e_m}|\boldsymbol{c}_{e_{-m}}) \end{aligned} \tag{2}$$

where

$$p(\boldsymbol{c}_{o_m}|\boldsymbol{\mu}) = \prod_{i=0}^{|\boldsymbol{c}_{o_m}|-1} p(c_{i,m}|\mu_{c_i}) \tag{3}$$

and

$$\begin{aligned} &p(\boldsymbol{w}_{e_m}|\boldsymbol{c}_{e_m}, \boldsymbol{w}_{e_{-m}}, \boldsymbol{z}_{e_m}) \\ &= \prod_{l=1}^{|c_{e_m}|} \left(\frac{\Gamma(n_{c_{e_m,l},-m}^{\cdot} + |V|\eta)}{\prod_w \Gamma(n_{c_{e_m,l},-m}^{w} + \eta)} \times \right. \\ &\quad \left. \frac{\prod_w \Gamma(n_{c_{e_m,l},-m}^{w} + n_{c_{e_m,l},m}^{w} + \eta)}{\Gamma(n_{c_{e_m,l},-m}^{\cdot} + n_{c_{e_m,l},m}^{\cdot} + |V|\eta)} \right) \end{aligned} \tag{4}$$

$\boldsymbol{c}_{e_{-m}}$ is the set of latent topics for all documents other than m, $\boldsymbol{z}_{e_m}$ is the assignment of the words in the m^{th} document to one of the latent topics, and $\boldsymbol{w}_{e_m}$ is the set of the words belonging to one of the latent topics in the the m^{th} document. $n_{c_{e_m,l},-m}^{w}$ is the number of instances of word w that have been assigned to the topic indexed by $c_{e_m,l}$, not including those in the document m.

Second, given the current state of the SSHLDA, we sample the $z_{m,n}$ variables of the underlying SSHLDA model as follows:

$$\begin{aligned} &p(z_{m,n} = j|\boldsymbol{z}_{-(m,n)}, \boldsymbol{w}, \boldsymbol{c}_m, \boldsymbol{\mu}) \\ &\quad \propto \frac{n_{-n,j}^{m} + \alpha}{n_{-n,\cdot}^{m} + |\boldsymbol{c}_m|} \cdot \frac{n_{-n,j}^{w_{m,n}} + \eta_{w_{m,n}}}{n_{-(m,n)}^{\cdot} + |V|} \end{aligned} \tag{5}$$

Having obtained the full conditional distribution, the Gibbs sampling algorithm is then straightforward. The $z_{m,n}$ variables are initialized to determine the initial state of the Markov chain. The chain is then run for a number of iterations, each time finding a new state by sampling each $z_{m,n}$ from the distribution specified by Equation (5). After obtaining individual word assignments $\mathbf{z}$, we can estimate the topic multinomials and the per-document mixing proportions. Specifically, the topic multinomials are estimated as:

$$\beta_{\mathbf{c_{m,j}},i} = p(w_i|z_{\mathbf{c_{m,j}}}) = \frac{\eta + n_{z_{\mathbf{c_{m,j}}}^{w_i}}}{|V|\eta + \sum n_{z_{\mathbf{c_{m,j}}}}^{\cdot}} \tag{6}$$

while the per-document mixing proportions fixed can be estimated as:

$$\theta_{m,j} = \frac{\alpha + n_{\cdot,j}^{m}}{|\mathbf{c_m}|\alpha + n_{\cdot,\cdot}^{m}}, j \in 1, ..., |\mathbf{c_m}| \tag{7}$$

5.1 Relation to Existing Models

In this section, we draw comparisons with the current state-of-the-art models for hierarchical topic

modeling (Blei et al., 2004; Petinot et al., 2011) and show that at certain choices of the parameters of our model, these methods fall out as special cases.

Our method generalises not only *hierarchical Latent Dirichlet Allocation* (hLDA), but also *Hierarchical Labeled Latent Dirichlet Allocation* (hLLDA). Our proposed model provides a unified framework allowing us to model hierarchical labels while to explore new latent topics.

Equivalence to hLDA As introduced in Section 2, hLDA is a unsupervised hierarchical topic model. In this case, there are no observed nodes, that is, the corpus has no hierarchical labels. This means $\boldsymbol{c}_m$ is equal to $\boldsymbol{c}_{e_m,m}$; meanwhile the factor $p(\boldsymbol{c}_{o_m,m}|\boldsymbol{\mu})$ is always equal to one because each document has root node, and this allows us to rewrite Formula (2) as:

$$p(\boldsymbol{c}_m|\boldsymbol{z},\boldsymbol{w},\boldsymbol{c}_{-m},\boldsymbol{\mu}) \propto p(\boldsymbol{w}_{c_m}|\boldsymbol{c},\boldsymbol{w}_{-m},\boldsymbol{z})p(\boldsymbol{c}_m|\boldsymbol{c}_{-m}) \quad (8)$$

which is exactly the same as the conditional distribution for $\boldsymbol{c}_m$, the L topics associated with document m in hLDA model. In this case, our model becomes equivalent to the hLDA model.

Equivalence to hLLDA hLLDA is a supervised hierarchical topic model, which means all nodes in hierarchy are observed. In this case, $\boldsymbol{c}_m$ is equal to $\boldsymbol{c}_{o_m,m}$, and this allows us to rewrite Formula (2) as:

$$p(\boldsymbol{c}_m|\boldsymbol{z},\boldsymbol{w},\boldsymbol{c}_{-m},\boldsymbol{\mu}) = p(\boldsymbol{c}_m|\boldsymbol{\mu}) \propto p(\boldsymbol{c}_{o_m}|\boldsymbol{\mu}) \quad (9)$$

which is exactly the same as the step " Draw a random path assignment c_m" in the generative process for hLLDA. Consequentially, in this sense our model is equivalent to hLLDA.

6 Experiments

We demonstrate the effectiveness of the proposed model on large, real-world datasets in the question answering and website category domains on two tasks: the topic modeling of documents, and the use of the generated topics for document clustering.

6.1 Datasets

To construct comprehensive datasets for our experiments, we crawled data from two websites. First, we crawled nearly all the questions and associated answer pairs (QA pairs) of two top categories of Yahoo! Answers: *Computers & Internet* and *Health*. This produced forty-three subcategories from 2005.11 to 2008.11, and an archive of 6,345,786 QA documents. We refer the Yahoo! Answer data as *Y_Ans*.

Table 2: The statistics of the datasets.

Datasets	#labels	#paths	Max level	#docs
Y_Ans	46	35	4	6,345,786
O_Hlth	6695	6505	10	54939
O_Home	2432	2364	9	24254

In addition, we first crawled two categories of Open Directory Project (ODP)*: *Home* and *Health*. Then, we removed all categories whose number of Web sites is less than 3. Finally, for each of Web sites in categories, we submited the url of each Web site to Google and used the words in the snippet and title of the first returned result to extend the summary of the Web site. We denote the data from the category *Home* as *O_Home*, and the data from the category *Health* as *O_Hlth*.

The statistics of all datasets are summarized in Table 2. From this table, we can see that these datasets are very diverse: *Y_Ans* has much fewer labels than *O_Hlth* and *O_Home*, but have much more documents for each label; meanwhile the depth of hierarchical tree for *O_Hlth* and *O_Home* can reach level 9 or above.

All experiments are based on the results of models with a burn-in of 10000 Gibbs sampling iterations, symmetric priors $\alpha = 0.1$ and free parameter $\eta = 1.0$; and for $\boldsymbol{\mu}$, we can obtain the estimation of μ_{c_i} by fixed-point iteration (Minka, 2003).

6.2 Case Study

With topic modeling, the top associated words of topics can be used as good descriptors for topics in a hierarchy (Blei et al., 2003; Blei and McAuliffe, 2010). We show in Figure 3 a pair of comparative example of the proposed model and a baseline model over *Y_Ans* dataset. The tree-based topic visualizations of Figure 3 (a) and (b) are the results of SSHLDA and Simp-hLDA.

We have three major observations from the example: (i) SSHLDA is a unified and generative model, after learning, it can obtain a hierarchy of topics;

*http://dmoz.org/

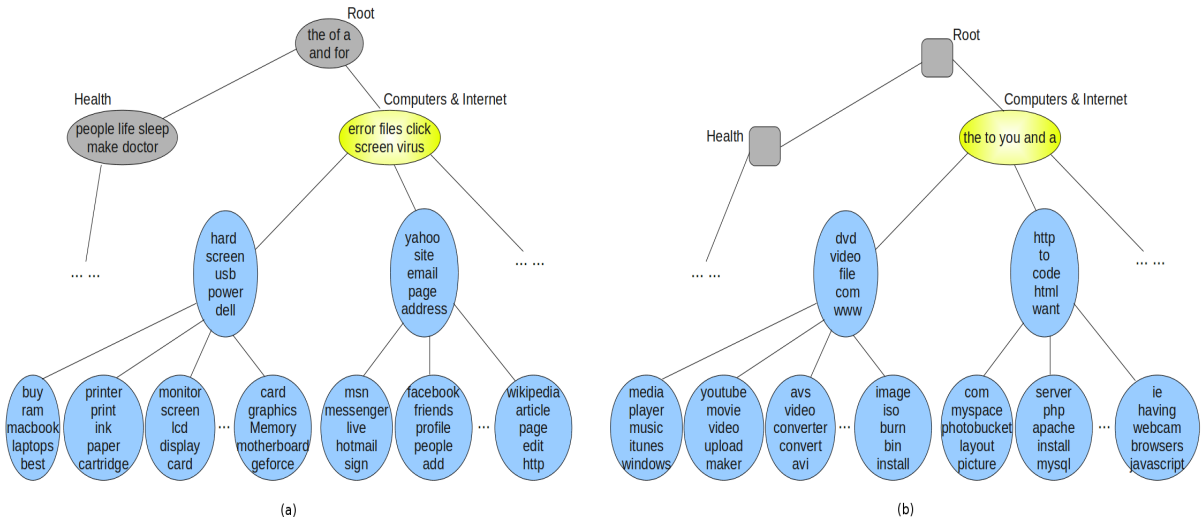

Figure 3: (a) A sub network discovered on *Y_Ans* dataset using SSHLDA, and the whole tree has 74 nodes; (b) A sub network discovered on *Y_Ans* dataset using Simp-hLDA algorithm, and the whole tree has 89 nodes. In both figures, the shaded and squared nodes are observed labels, not topics; the shaded and round nodes are topics with observed labels; blue nodes are topics but without labels and the yellow node is one of leaves in hierarchy of labels. Each topic represented by top 5 terms.

while Simp-hLDA is a heuristic method, and its result is a mixture of label nodes and topical nodes. For example, Figure 3 (b) shows that the hierarchy includes label nodes and topic nodes, and each of labeled nodes just has a label, but label nodes in Figure 3 (a) have their corresponding topics. (ii) During obtaining a hierarchy, SSHLDA makes use of the information from observed labels, thus it can generate a logical, structual hierarchy with parent-child relations; while Simp-hLDA does not incorporate prior information of labels into its generation process, thus although it can obtain a hierarchy, many parent-child pairs have not parent-child relation. For example, in Figure 3 (b), although label "root" is a parent of label "Computers & Internet", the topical words of label "Computers & Internet" show the topical node is not a child of label "root". However, in Figure 3 (a), label "root" and "Computers & Internet" has corresponding parent-child relation between their topical words. (iii) In a hierarchy of topics, if a topical node has corresponding label, the label can help people understand descendant topical nodes. For example, when we know node "error files click screen virus" in Figure 3 (a) has its label "Computers & Internet", we can understand the child topic "hard screen usb power dell" is about "computer hardware". However, in Figure 3 (b), the labels in parent nodes cannot provide much information to understand descendant topical nodes because many label nodes have not corresponding right topical words, such as label "Computers & Internet", its topical words, "the to you and a", do not reflect the connotation of the label.

These observations further confirm that SSHLDA is better than the baseline model.

6.3 Perplexity Comparison

A good topic model should be able to generalize to unseen data. To measure the prediction ability of our model and baselines, we compute the perplexity for each document d in the test sets. *Perplexity*, which is widely used in the language modeling and topic modeling community, is equivalent algebraically to the inverse of the geometric mean per-word likelihood (Blei et al., 2003). Lower perplexity scores mean better. Our model, SSHLDA, will compare with three state-of-the-art models, i.e. Simp-hLDA, hLDA and hLLDA. Simp-hLDA has been introduced in Section 1, and hLDA and hLLDA has been reviewed in Section 2. We keep 80% of the data collection as the training set and use the remaining collection as the held-out test set. We build the mod-

els based on the train set and compute the preplexity of the test set to evaluate the models. Thus, our goal is to achieve lower perplexity score on a held-out test set. The perplexity of M test documents is calculated as:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\sum_{m=1}^{N_d}\log p(w_{dm})}{\sum_{d=1}^{M} N_d}\right\} \quad (10)$$

where D_{test} is the test collection of M documents, N_d is document length of document d and w_{dm} is m^{th} word in document d.

We present the results over the *O_Hlth* dataset in Figure 4. We choose top 3-level labels as observed, and assume other labels are not observed, i.e. $l = 3$. From the figure, we can see that the perplexities of SSHLDA, are lower than that of Simp-hLDA, hLDA and hLLDA at different value of the tree height parameter, i.e. $L \in \{5, 6, 7, 8\}$. It shows that the performance of SSHLDA is always better than the state-of-the-art baselines, and means that our proposed model can model the hierarchical labeled data better than the state-of-the-art models. We can also obtain similar experimental results over *Y_Ans* and *O_Home* datasets, and their detailed description is not included in this paper due to the limitation of space.

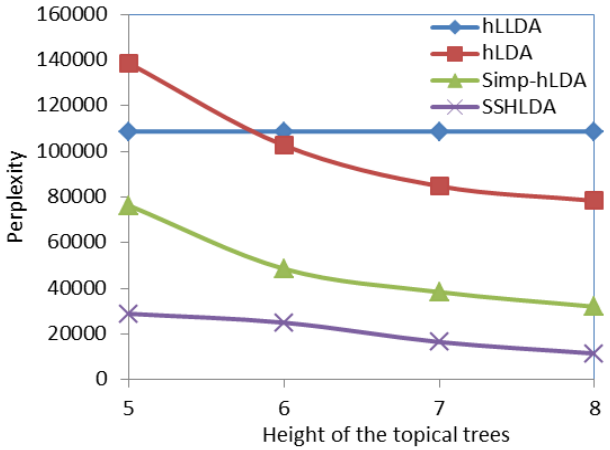

Figure 4: Perplexities of hLLDA, hLDA, Simp-hLDA and SSHLDA. The results are run over the O_Hlth dataset, with the height of the hierarchy of observed labels $l = 3$. The X-axis is the height of the whole topical tree (L), and Y-axis is the perplexity.

6.4 Clustering performance

To evaluate indirectly the performance of the proposed model, we compare the clustering performance of following systems: 1) the proposed model; 2) Simp-hLDA; 3) hLDA; 4) agglomerative clustering algorithm. There are many agglomerative clustering algorithms, and in this paper, we make use of the single-linkage method in a software package called CLUTO (Karypis, 2005) to obtain hierarchies of clusters over our datasets, with words as features. We refer the method as *h-clustering*.

Given a document collection *DS* with a *H-level* hierarchy of labels, each label in the hierarchy and corresponding documents will be taken as the ground truth of clustering algorithms. The hierarchy of labels denoted as *GT-tree*. The process of evaluation is as follows. First, we choose top l-$level$ labels in GT-tree as an observed hierarchy, i.e. Base Tree (BT), and we need to construct a L-$level$ hierarchy ($l < L <= H$) over the documents DS using a model. The remaining labels in GT-tree and corresponding documents are the ground truth classes, each class denoted as C_i. Then, (i) for h-clustering, we run single-linkage method over the documents DS. (ii) for Simp-hLDA, hLDA runs on the documents in each leaf-node in BT, and the height parameter is $(L - l)$ for each hLDA. After training, each document is assigned to top-1 topic according to the distribution over topics for the document. Each topic and corresponding documents forms a new cluster. (iii) for hLDA, hLDA runs on all documents in DS, and the height parameter is L. Similar to Simp-hLDA, each document is assigned to top-1 topic. Each topic and corresponding documents forms a new cluster. (iv) for SSHLDA, we set height parameter as L. After training, each document is also assigned to top-1 topic. Topics and their corresponding documents form a hierarchy of clusters.

6.4.1 Evaluation Metrics

For each dataset we obtain corresponding clusters using the various models described in previous sections. Thus we can use clustering metrics to measure the quality of various algorithms by using a measure that takes into account the overall set of clusters that are represented in the new generated part of a hierarchical tree.

One such measure is the FScore measure, intro-

duced by (Manning et al., 2008). Given a particular class C_r of size n_r and a particular cluster S_i of size n_i, suppose n_{ri} documents in the cluster S_i belong to C_r, then the FScore of this class and cluster is defined to be

$$F(C_r, S_i) = \frac{2 \times R(C_r, S_i) \times P(C_r, S_i)}{R(C_r, S_i) + P(C_r, S_i)} \quad (11)$$

where $R(C_r, S_i)$ is the recall value defined as n_{ri}/n_r, and $P(C_r, S_i)$ is the precision value defined as n_{ri}/n_i for the class C_r and the cluster S_i. The FScore of the class C_r, is the maximum FScore value attained at any node in the hierarchical clustering tree T. That is,

$$F(C_r) = \max_{S_i \in T} F(C_r, S_i). \quad (12)$$

The FScore of the entire clustering solution is then defined to be the sum of the individual class FScore weighted according to the class size.

$$FScore = \sum_{r=1}^{c} \frac{n_r}{n} F(C_r), \quad (13)$$

where c is the total number of classes. In general, the higher the FScore values, the better the clustering solution is.

6.4.2 Experimental Results

Each of hLDA, Simp-hLDA and SSHLDA needs a parameter—the height of the topical tree, i.e. L; and for Simp-hLDA and SSHLDA, they need another parameter—the height of the hierarchical observed labels, i.e l. The h-clustering does not have any height parameters, thus its FScore will keep the same values at different height of the topical tree. With choosing the height of hierarchical labels for *O_Home* as 4, i.e. $l = 4$, the results of our model and baselines with respect to the height of a hierarchy are shown in Figure 5.

From the figure, we can see that our proposed model can achieve consistent improvement over the baseline models at different height, i.e. $L \in \{5, 6, 7, 8\}$. For example, the performance of SSHLDA can reach 0.396 at height 5 while the h-clustering, hLDA and hLLDA only achieve 0.295, 0.328 and 0.349 at the same height. The result shows that our model can achieve about 34.2%, 20.7% and 13.5% improvements over h-clustering, hLDA and hLLDA at height 5. The improvements are significant by t-test at the 95% significance level. We can also obtain similar experimental results over *Y_Ans* and *O_Hlth*. However, for the same reason of limitation of space, their detailed descriptions are skipped in this paper.

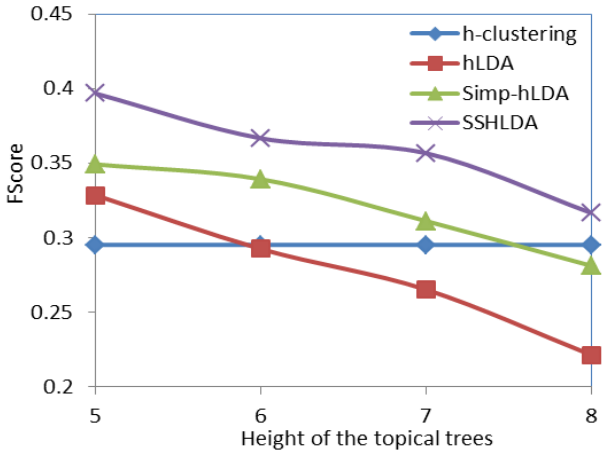

Figure 5: FScore measures of h-clustering, hLDA, Simp-hLDA and SSHLDA. The results are run over the O_Home dataset, with the height of the hierarchy of observed labels $l = 3$. The X-axis is the height of the whole topical tree (L), and Y-axis is the FScore measure.

7 Conclusion and Future work

In this paper, we have proposed a semi-supervised hierarchical topic models, i.e. SSHLDA, which aims to solve the drawbacks of hLDA and hLLDA while combine their merits. Specially, SSHLDA incorporates the information of labels into generative process of topic modeling while exploring latent topics in data space. In addition, we have also proved that hLDA and hLLDA are special cases of SSHLDA. We have conducted experiments on the Yahoo! Answers and ODP datasets, and assessed the performance in terms of Perplexity and FScore measure. The experimental results show that the prediction ability of SSHLDA is the best, and SSHLDA can also achieve significant improvement over the baselines on Fscore measure.

In the future, we will continue to explore novel topic models for hierarchical labeled data to further improve the effectiveness; meanwhile we will also apply SSHLDA to other media forms, such as image, to solve related problems in these areas.

Acknowledgments

This work was partially supported by NSFC with Grant No.61073082, 60933004, 70903008 and NExT Search Centre, which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

References

D. Blei and J. Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems*, 18:147.

D.M. Blei and J.D. McAuliffe. 2007. Supervised topic models. In *Proceeding of the Neural Information Processing Systems(nips)*.

D.M. Blei and J.D. McAuliffe. 2010. Supervised topic models. *Arxiv preprint arXiv:1003.0783*.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:106.

C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers. 2008a. Modeling documents by combining semantic concepts with unsupervised statistical learning. *The Semantic Web-ISWC 2008*, pages 229–244.

C. Chemudugunta, P. Smyth, and M. Steyvers. 2008b. Combining concept hierarchies and statistical topic models. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1469–1470. ACM.

C. Chemudugunta, P. Smyth, and M. Steyvers. 2008c. Text modeling using unsupervised topic models and concept hierarchies. *Arxiv preprint arXiv:0808.0973*.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

T. Hofmann. 1999. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, page 21. Citeseer.

G. Karypis. 2005. Cluto: Software for clustering high dimensional datasets. *Internet Website (last accessed, June 2008), http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview*.

S. Lacoste-Julien, F. Sha, and M.I. Jordan. 2008. ndisclda: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems*, 21.

W. Li and A. McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

C.D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

D. Mimno, W. Li, and A. McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pages 633–640. ACM.

Z.Y. Ming, K. Wang, and T.S. Chua. 2010. Prototype hierarchy based clustering for the categorization and navigation of web collections. In *Proceeding of the 33rd international ACM SIGIR*, pages 2–9. ACM.

T.P. Minka. 2003. Estimating a dirichlet distribution. *Annals of Physics*, 2000(8):1–13.

A. Perotte, N. Bartlett, N. Elhadad, and F. Wood. 2011. Hierarchically supervised latent dirichlet allocation. *Neural Information Processing Systems (to appear)*.

Y. Petinot, K. McKeown, and K. Thadani. 2011. A hierarchical model of web summaries. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies: short papers-Volume 2*, pages 670–675. ACL.

D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. 2009a. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

D. Ramage, P. Heymann, C.D. Manning, and H. Garcia-Molina. 2009b. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63. ACM.

D. Ramage, C.D. Manning, and S. Dumais. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465. ACM.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

T.N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. 2011. Statistical topic models for multi-label document classification. *Arxiv preprint arXiv:1107.2462*.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.