Received November 24, 2019, accepted December 15, 2019, date of publication December 18, 2019, date of current version December 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2960468

A Novel Hierarchical Topic Model for Horizontal Topic Expansion With Observed Label Information

XI ZOU, YUELONG ZHU, JUN FENG, JIAMIN LU, AND XIAODONG LI
School of Computer and Information, Hohai University, Nanjing 211100, China

Corresponding author: Jiamin Lu (jiamin.luu@hhu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0407901 and Grant 2017YFC0405806, in part by the Fundamental Research Funds for the Central Universities under Grant 2019B64614 and Grant 2019B15514, in part by the National Natural Science Foundation of China under Grant 61602149, Grant 61370091, and Grant 61602151, and in part by the Jiangsu Collaborative Innovation Center for Cultural Creativity under Grant XYN1702.

ABSTRACT Hierarchical topic models, such as hierarchical Latent Dirichlet Allocation (hLDA)and its variations, can organize topics into a hierarchy automatically. On the other hand, there are lots of documents associated with hierarchical label information. Incorporating these information into the topic modeling process can help users to obtain a more reasonable hierarchical structure. However, after analyzing various real-world datasets, we find that these hierarchical labels are ambiguous and conflicting in some levels, which introduces error and restriction to the latent topic and the hierarchical structure exploration process. We call it the horizontal topic expansion problem. To address this problem, in this paper, we propose a novel hierarchical topic model named horizontal and vertical hierarchical topic model (HV-HTM), which aims to incorporate the observed hierarchical label information into the topic generation process, while keeping the flexibility of horizontal and vertical expansion of the hierarchical structure in the modeling process. We conduct experiments on BBC news and Yahoo! Answers datasets and evaluate the effectiveness of HV-HTM on three evaluation metrics. The experimental results show that HV-HTM has a significant improvement on topic modeling, compared to the state-of-the-art models, and it can also obtain a more interpretable hierarchical structure.

INDEX TERMS Topic modeling, hierarchical topic model, hierarchical latent Dirichlet allocation, label information.

I. INTRODUCTION

Topic modeling is one of the most popular research areas in Natural Language Process (NLP), which aims at digging out the latent topics from a large collection of documents. Topic models, such as Latent Dirichlet allocation (LDA) [1], have been proven to be useful in extracting latent topics. As a generative model, each document is viewed as a mixture of topics, and the topic is viewed as a mixture of words. However, the topics of LDA model is "flat" without considering the hierarchical relationship among the topics, such as the parent-child and the sibling relationships. Therefore, hierarchical topic models, like hierarchical Latent Dirichlet Allocation (hLDA) [2], are proposed to relax this restriction. Those models make use of the Chinese restaurant process and its extension for constructing topic hierarchy. On the other hand, there are lots of documents are manually organized in hierarchical directories or labeled hierarchically, such as question answering and news categories. Topic models which incorporate these hierarchical label information into the modeling process are attracting more and more attention. The hierarchical label information is treated as observed nodes of the topical tree to supervise the generative process of the hierarchical structure. Therefore, these methods usually obtain a better performance than traditional flat topic models.

However, most hierarchical topic models concentrate on generating specific child nodes in deeper vertical level. The label information in those methods are strong constraint, which means once a node has observed child nodes, it can not generate other new child nodes. As a result, these methods do not have the ability to expand the topical tree horizontally. On the other hand, there are some hierarchical label

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

information associated with documents are ambiguous and conflicting. For instance, after analyzing Yahoo! Answers dataset carefully, which labels each question & answer pair with hierarchical category, we have two main discoveries: (1) the sub-categories of some categories are conflicting. For example, ''France'', ''United Kingdom'', and ''Europe'' are sub-categories of categories ''Travel''. (2) some categories are represented with ''Other'', such as ''Other Diseases'' and ''Other General Health Care''. If topic models employ these hierarchical label information to control the generation of child nodes, a lot of inaccurate supervision is introduced. We call this problem the horizontal topic expansion problem.

To address this problem, we focus on hierarchical topic models incorporating observed hierarchical label information and how to expand the topical tree horizontally and vertically. In this paper, we present the novel horizontal and vertical hierarchical topic model, called HV-HTM. This model first select a labeled node or generate a new node in an observed hierarchy through a general version of the Chinese restaurant process, which incorporates label information into the topic generation process and keeps the flexibility of the horizontal topic expansion. Then this model generates a sub-hierarchy in deeper vertical level. On the basis of this idea, HV-HTM can flexibly expand topical tree both horizontally and vertically, and generate a more reasonable hierarchical structure.

We demonstrate the effectiveness of the proposed model in two large, real-world datasets. The case study shows that our HV-HTM model can address the horizontal expansion problem and generate a clear and readable hierarchical structure. The results on three metrics verify that HV-HTM outperforms the state-of-the-art hierarchical topic models. The main contribution of this paper is threefold. (1) We extend the Chinese restaurant process to a general version, and present three strategies to identify the number of customers in the occupied table. In this way, observed label information is incorporated into topic generation process effectively. (2) We propose a horizontal and vertical hierarchical topic model (HV-HTM), enabling the topical tree to expand horizontally and overcoming the defects of the observed hierarchical label information. (3) We develop a Gibbs sampling algorithm for the proposed model to estimate proper model parameters. We conduct extensive experiments on large datasets and evaluate the performance of our model.

The remainder of this paper is organized as follows. In section 2, a brief review of the related works are illustrated, and some preliminaries which is essential for understanding this paper are introduced in Section 3. Then, we propose our model and detail a parameter inference algorithm in Section 4. Section 5 presents experimental result and the comparison with other models. Finally, we conclude the paper and outline future work in section 6.

II. RELATED WORKS

In this section, we give a brief review of the related works from three aspects. The first is topic modeling, the research topic of this paper. The Second is hierarchical topic model, the problem solved in this paper. The third is incorporating label information into topic modeling process, the solution of this paper.

A. TOPIC MODELING

Latent sematic Analysis (LSA) [3] can be considered as the earliest attempt of topic modeling, although there is no explicit topic concept in LSA. Hofmann [4] presents a proper probabilistic generative model named Probabilistic Latent Semantic Analysis (PLSA), in which each document is a mixture of topics, and each topic is a distribution of vocabulary. Similar to PLSA, Blei *et al.* [5] propose Latent Dirichlet Allocation (LDA), except that topic parameters in LDA are assumed to have Dirichlet priors, which makes LDA is effective. Since then, the researchers have proposed various models based on LDA. Topic over Time (ToT) [6] and Dynamic Topic Model (DTM) [7] are introduced to obtain the evolution of topics over time in a sequentially organized corpus. Correlated Topic Model (CTM) [8] can represent pairwise topic correlations. Author-Topic Model (ATM) [9] provides a distribution of topics for each author to find relationships among authors, topics, words and documents. Zhao *et al.* [10] propose Twitter-LDA, which aims to mine topics from short texts such as tweets. However, the topic models mentioned above, such as LDA, are flat, with no direct hierarchical relationship among topics. Therefore, these models are suitable for revealing the potential topics of corpus, they fail to indicate the level of the topics.

B. HIERARCHICAL TOPIC MODEL

In order to capture the topic hierarchies from textual data, many researchers have extended traditional topic models to obtain hierarchical information on the topics over the past decade. The model of hierarchical Latent Dirichlet Allocation (hLDA) [11] regards topic hierarchies as random variables. Moreover, a stochastic process called nested Chinese restaurant process (nCRP) [12] is used as a prior distribution to model topic hierarchy into a L-level tree. Then a document is generated by choosing a path from the root node to a leaf node, repeatedly sampling topics along the path, and sampling the words from the selected topics. The biggest advantage of hLDA is that nCRP expresses the uncertainty of the L-level trees rather than assuming a fixed tree structure. On the basis of hLDA, some variations are proposed. Those variations can be divided into the following four categories: (1) to explore new probabilistic modeling framework, (2) to fuse additional information aspect, (3) to organize structure with prior knowledge, and (4) to incorporate label information of documents.

Pachinko Allocation Model (PAM) [13] is a case of exploring new probabilistic modeling framework. The correlation of topics in PAM is modelled by a direct acyclic graph (DAG) instead of a tree in hLDA. Mimno *et al.* [14] extend PAM to a hierarchical version called Hierarchical PAM (HPAM), which enables documents to have multiple parent topics. Kim *et al.* [15] propose hierarchical aspect sentiment model (HASM) is a variation of fusing additional information aspect. Each node in HASM is a two-level tree, whose root represents an aspect and the children represent sentiment associated with it. Zhu *et al.* [16] also propose a hierarchical opinion phrase (HOP) model, which assumes that the assignment of the viewpoint topics follows two nested Chinese restaurant process. Guided hierarchical topic model (GHTM) [17] allocates the prior knowledge to the Dirichlet Forest prior is an attempt to organize structure by domain knowledge. Yu *et al.* [18] propose twitter hierarchical Latent Dirichlet Allocation (thLDA), which uses word2vec to analyze the semantic relationships of words in tweets to obtain a more effective dimension. Finally, to incorporate label information of document, hierarchical Labeled Latent Dirichlet Allocation (hLLDA) [19] uses the hierarchy of the DMOZ. This hierarchy provides a backbone around which the model crystalizes hierarchical topic model. Our model falls into this category, and we will review more relevant works in next subsection.

Besides the Bayesian parametric models, there are several studies that focus on nonparametric hierarchical topic modeling, which do not require setting the number of topics in advance. Ahmed *et al.* [20] introduce the nested Chinese restaurant franchise process to combine the advantages of the hierarchical Dirichlet process (HDP) and the nested Chinese restaurant process. Lim *et al.* [21] propose Hierarchical Pitman-Yor process (HPYP), which is a simple network of Pitman-Yor process (PYP) nodes since all distributions on the probability vectors are modelled by the PYP.

C. INCORPORATING LABEL INFORMATION

This paper will concentrate on hierarchical topic models that take account of label information of documents to supervise topic generation. Those extensions belong to supervised or semi-supervised hierarchical topic models. In addition to the hLLDA mentioned above, Perotte *et al.* [22] propose hierarchically supervised latent Dirichlet allocation (HSLDA), which jointly models the hierarchy of labels and topics. But a label in HSLDA is not associated with a probability distribution. Nguyen *et al.* [23] introduce supervised hierarchical latent Dirichlet allocation (SHLDA), which extends the nested Chinese restaurant process to allows documents to have multiple paths through the tree. On the other hand, some semi-supervised models are also proposed, which aim to detect latent topics automatically in the data space while incorporating the information from the observed hierarchical labels into the modeling process. Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA) [24] defines two concepts: labeled topic which refers the topic with a corresponding label, and latent topic which is unseen and without a label. The generative process is nearly same as hLDA, except adding a condition in path generation. When generating the path in each level, if nodes in this level have been observed, directly sampling a node from a multinomial distribution. Constrained-hLDA [25] first extracts path-constraints to pre-establish a part of the infinite tree structure, then extends the nCRP to integrate path constraint. Generalized SSHLDA (G-SSHLDA) [26] can insert a latent subtree at any level in an observed hierarchy. In the latent subtree, its root can be an arbitrary observed node, and the other nodes are latent nodes.

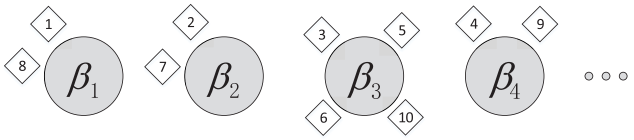


FIGURE 1. **A partition of ten customers by Chinese restaurant process (CRP). Each grey circle indicates a table in the restaurant, each square represents a customer. In this example, ten customers occupy four of the infinite number of tables.**

Our work in this paper is inspired by the approaches reviewed above, especially hLDA and SSHLDA. We concentrate on obtaining topic hierarchies while incorporating hierarchical labels. While SSHLDA generates latent topics starting from observed leaf nodes. Our model proposed in this paper is able to generate sibling nodes of observed nodes from level 2, which achieves the horizontal expansion of the topics in each level.

III. PRELIMINARIES

In this section, we first briefly describe the fundamental of this study, including the Chinese restaurant process, the nested Chinese restaurant process and hierarchical Latent Dirichlet Allocation. Then the notations used in this paper are defined.

A. CHINESE RESTAURANT PROCESS

The Chinese restaurant process (CRP) [27] is a discrete-time stochastic process which generates a probability distribution on partitions of integers. The CRP can be defined by imagining the following metaphor. Suppose there is a Chinese restaurant with an infinite number of tables, and each table has infinite capacity. The sequence of N customers $\{1, 2, \ldots, N\}$ come to the restaurant. The first customer sits at the first table; and the nth subsequent customer can sit at the ith occupied table or choose an unoccupied new table drawn from the following distribution:

$$\begin{aligned} p(table\ i \mid previous\ customer) &= \frac{n_i}{\gamma + n - 1} \\ p(new\ table \mid previous\ customer) &= \frac{\gamma}{\gamma + n - 1} \end{aligned} \tag{1}$$

where n_i is the number of customers sitting at table i, $n-1$ is the total number of customers before the nth customer arrives, and γ is a parameter, which aims to control the probability of the customer chooses a new table. After N customers have sat down, their seating plan represents a partition of N customers as illustrated in Fig. 1.

Therefore, each table can be treated as a mixture component and each customer can be treated as data, that the CRP can be exploited to associate data with a component. Meanwhile, the CRP have the advantages to represent the

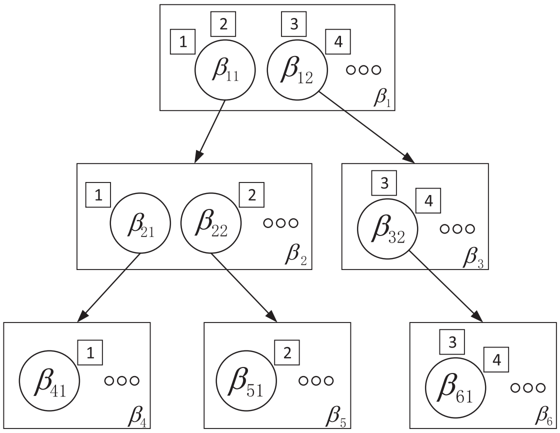


FIGURE 2. The paths of four tourists with three-day travels following the nest Chinese restaurant process (nCRP). Each rectangle represents a restaurant tables, each circle represents a table which points to anther restaurant.

uncertainty over the number of mixture components in a mixture model.

B. NESTED CHINESE RESTAURANT PROCESS

The nested Chinese restaurant process (nCRP) is an variation of standard CRP, can be described by imagining the following scenario. Suppose in a city there are an infinite number of restaurants, each of which has an infinite number of tables. One restaurant is identified as the root restaurant, and there is a card on each of its infinite tables with a name of anther restaurant. In other words, the root restaurant and the restaurants referred to on its tables' cards are organized into a 2-level tree. Similarly, each table in child restaurants has cards referring to other restaurants, and this structure repeats infinitely many times. Consequently, all restaurants in the city form an infinitely-branched tree. When a tourist arrives at this city, he comes into the root restaurant and selects a table according to equation (1) on the first day. Next day, he comes to the restaurant indicated by the card on the first day's tables and chooses a table, again from equation (1). After L-day travel, he has sat at L restaurants, which constitute a path from root restaurant to a restaurant at the Lth level. For all M tourists, the collection of paths, which followed by each tourist, describe a L-level subtree of the infinite tree described above. Fig. 2 is an example of such a tree.

As an extension of CRP, each restaurant (node) in the tree can be regarded as a topic, each tourist can be regraded as a document, and the path corresponding to each tourist can be regarded as the topics associated with a document. Thus, the nCRP can be used to model topic hierarchies and express the uncertainty about the hierarchical structure.

C. HIERARCHICAL LATENT DIRICHLET ALLOCATION

Hierarchical Latent Dirichlet Allocation (hLDA) treats topic hierarchies as random variables, and applies the nCRP as a prior distribution to organize topic hierarchy into a L-level tree rather than a flat structure. Specifically, a certain document is generated by first choosing a path from the root node to a leaf node by the nCRP, and then drawing a topic proportion vector from Dirichlet distribution. Secondly, for each word in the document, hLDA repeatedly samples topics according to the topic proportion, and samples each word of the documents from a corresponding multinomial distribution of the selected topic. In this way, hLDA obtains topic hierarchies as well as topic probability distribution across vocabulary simultaneously after a certain number of iterations.

The process of hLDA can be understood as to classify a collection of documents according to the co-occurrence of words within documents. A set of function words (e.g., "a" and "the") share in all documents. Thus the root restaurant is represented by a set of function words. At each subsequent level, in order to divide the documents into several categories, hLDA tries to find a set of more specific words in those documents. Consequently, hLDA obtains topic hierarchies in which more abstract topics are near the root and more concrete topics are near the leaves. As a unsupervised model, hLDA is a pure data-driven approach without any extra information about the corpus. In addition, topic hierarchies in hLDA are unfixed, which means that the number of topics in hLDA can grow dynamically as the growth of the corpus.

TABLE 1. Notations used throughout this paper.

Symbol	Description
M	The number of documents in the corpus
N_m	The number of words of document m
V	The number of tokens in the vocabulary
T	The collection of paths generated by nCRP
α	Parameter for Dirichlet distribution over θ_m
η	Parameter for Dirichlet distribution over β_k
γ	Parameter for the CRP
$\boldsymbol{\delta}$	The number of local customers before tourists arrive in the restaurant
θ_m	Distribution of topic over document m
β_k	Distribution of vocabulary over topic k
$\boldsymbol{c}_m$	The path in the tree for document m
$\boldsymbol{c}_{-m}$	The paths for all documents except document m
$c_{m,l}$	The lth level of path c_m
$\boldsymbol{z}_m$	The topic allocation of every word in document m
$\boldsymbol{z}_{-m}$	The topic allocation except document m
$z_{m,n}$	The topic allocation of word n of document m
$z_{-(m,n)}$	The topic allocation leaving out word n of document m
$\boldsymbol{w}_m$	The words of documents m
$\boldsymbol{w}_{-m}$	The words except document m
$w_{m,n}$	The nth word of document m
$n_{c_{m,l}}^{(\cdot)}$	The assigned to topic indexed by $c_{m,l}$
$n_{c_{m,l}}^{(w)}$	The number of word w which is assigned to topic indexed by $c_{m,l}$

D. NOTATION

The notations used throughout this paper is listed in Table 1. Note that, the subscript "$-$" indicates the following elements is removed. When an index is replaced with "$\cdot$", it represents the summation of all possible choices of the index.

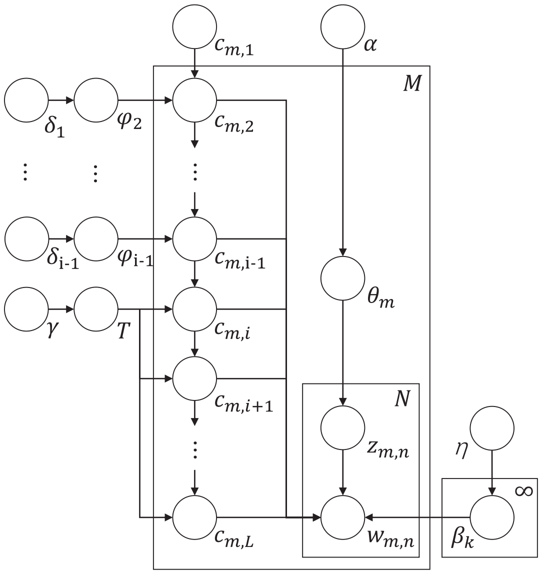


FIGURE 3. The graphical model of HV-HTM. Boxes are plate notations representing repetition.

IV. THE HORIZONTAL AND VERTICAL HIERARCHICAL TOPIC MODEL

In this section, we propose the Horizontal and Vertical Hierarchical Topic model (HV-HTM), a novel probabilistic graphical model that explores latent topics and topic hierarchies simultaneously. The generative process of the paths and topics in each document is described firstly. Then a Gibbs sampling algorithm is developed to infer the model parameters.

A. GENERATIVE PROCESS

Our HV-HTM model incorporates hierarchical label information among documents into the modeling process. In addition, it is capable to expand topic hierarchies in horizontal and vertical flexibly. Fig. 3 shows the probability graphical model of HV-HTM. Compared with SSHLDA, HV-HTM integrates the CRP with observed label information to improve the path sampling, when there are observed nodes in this level. The generative process of HV-HTM consists of two steps: sampling the per-document path c_m for each document, sampling the topic allocation $z_{m,n}$ for each word in the document.

Specifically, we extend the CRP to a more general situation, named gCRP. Suppose there are several tables in the restaurant that have been occupied by local customers before the tourists arrive. When the first tourist enters into this restaurant, he needs to decide whether to sit at a new table or select an occupied table. The subsequent nth ($n \in \{2, \dots, N\}$) tourist also needs to consider the local customers and previous $n-1$ tourists when he selects a table. Thus the probability distribution for nth tourist to choose ith occupied table or a new table can be described as follows:

$$\begin{aligned} p(table\ i) &= \frac{n_i + \delta_i}{\gamma + \delta + n - 1} \\ p(new\ table) &= \frac{\gamma}{\gamma + \delta + n - 1} \end{aligned} \tag{2}$$

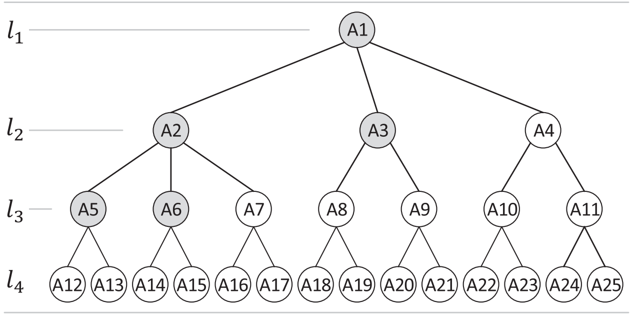


FIGURE 4. An example of HV-HTM, where the level of the topical tree is 4. The shaded nodes indicate labeled topics corresponding to the observed label information. The circled nodes are latent topics generated automatically by HV-HTM model. After learning, each node in the tree is associated with a probability distribution over vocabulary.

where δ_i is the number of local customers in table i, n_i is the number of tourists sitting on table i before nth tourist arrives, δ is the total number of local customers in the restaurant before tourists arrive, $\delta + n - 1$ represents the total number of customers before the nth tourist arrives. In this way, we can relate the observed label information in each level to the occupied tables in each restaurant. And these observed labels will guide the the generative process of the path. Now, there is a problem is to determine how many local customers in the occupied table. We present three strategies to identify the number of local customers in the occupied table, as described below:

- directly set to 1, which means this strategy only concerns about whether a table is occupied by local customers. When observed label information is not related with corresponding documents, this setting is still workable. We name our model with this strategy as HV-HTM(1).
- set to the number of documents corresponding to each label. This setting completely reflect original document allocation corresponding to every label, but this strategy will cause the probability of selecting a new table is lower. We name it as HV-HTM(num).
- set to a new value according to the proportion of observed labels in each level. By iteratively sampling label from corresponding multinomial distribution until each label is sampled, and counting the number of samples for each label. Those values then are set to the the number of local customers in the occupied table. This strategy avoids the lower chance to select a new table, as well as keeping the distribution of original document allocation corresponding to each label. We name it as HV-HTM(sample).

As an example shown in Fig. 4, there are a collection of documents with hierarchical label information: $\{A1, A2, A3, A5, A6\}$. Assuming the height of the desired topical tree is $L = 4$. All circled nodes are latent topics, which are generated automatically by HV-HTM model. It is notable that the latent topics can be generated even if there are labeled children in this level. For example, although node $A1$

has labeled children nodes $\{A2, A3\}$, HV-HTM model can generate latent topic $A4$ in this level. Thus, a possible path for a document m can be: It starts from root node $A1$, then chooses node $A4$ at level 2, and chooses $A10, A22$ respectively in the following level. Finally, the model obtains a path $\boldsymbol{c}_m = \{A1, A4, A10, A22\}$ for document m. When M documents in the corpus obtain M topic paths, a topical tree is constructed. After obtain per-document path c_m, topic for each word in the document is then sampled from this path.

B. INFERENCE AND PARAMETER ESTIMATION

Given the collection of documents, the words $\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$ is observed, and the label information of the corpus is also given. Our goal is to infer the hidden variables $\boldsymbol{c}$ and $\boldsymbol{z}$, which maximize the posterior probability distribution $p(\boldsymbol{c}, z|\boldsymbol{w})$. Since exact inference of the posterior distribution is intractable. Gibbs sampling [28] algorithm is employed to approximate the hidden variables. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations to approximate a complex multivariate probability distribution. In the HV-HTM model, the posterior probability distribution is shown in equation (3).

$$p(\boldsymbol{c}, z|\boldsymbol{w}, \alpha, \eta, \gamma, \boldsymbol{\delta}) = p(\boldsymbol{c}|\boldsymbol{w}, \gamma, \boldsymbol{\delta}, \eta) \times p(z|\boldsymbol{w}, \boldsymbol{c}, \alpha, \eta) \tag{3}$$

where α and η are concentration parameters of Dirichlet distribution, the smaller the value, the fewer components have high probability in each single sample. γ and is parameter of CRP, a larger value will lead to a higher probability that a tourist selects a new table. $\boldsymbol{\delta}$ is the number of local customers before tourists arrive in the restaurant. In order to exploit Gibbs sampling algorithm, we need to calculate the per-document conditional probability, which is shown in equation (4).

$$\begin{aligned} &p(\boldsymbol{c}_m, z_m|\boldsymbol{w}, \boldsymbol{c}_{-m}, z_{-m}, \alpha, \eta, \gamma, \boldsymbol{\delta}) \\ &\quad = p(\boldsymbol{c}_m|\boldsymbol{w}, \boldsymbol{c}_{-m}, z, \gamma, \boldsymbol{\delta}, \eta) \times p(z_m|\boldsymbol{w}, \boldsymbol{c}, z_{-m}, \alpha, \eta) \end{aligned} \tag{4}$$

This conditional probability indicates the generation probability of document m with given other document paths and topic allocations. According to (4), the process of Gibbs sampling algorithm can be divide into two separated parts. First, we randomly sample path $\boldsymbol{c}_m$ according to $p(\boldsymbol{c}_m|\boldsymbol{w}, \boldsymbol{c}_{-m}, z, \gamma, \boldsymbol{\delta}, \eta)$. Then we repeatedly sample word-wise topic allocation $z_{m,n}$ according to $p(z_{m,n}|\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{z}_{-(m,n)}, \alpha, \eta)$.

1) PATH SAMPLING

For the path sampling, since the words in the document are observed, the conditional probability distribution $p(\boldsymbol{c}_m|\boldsymbol{w}, \boldsymbol{c}_{-m}, z, \gamma, \boldsymbol{\delta}, \eta)$ can be expressed as follows:

$$\begin{aligned} &p(\boldsymbol{c}_m|\boldsymbol{w}, \boldsymbol{c}_{-m}, z, \gamma, \boldsymbol{\delta}, \eta) \\ &\qquad \propto p(\boldsymbol{c}_m, \boldsymbol{w}_m|\boldsymbol{w}_{-m}, \boldsymbol{c}_{-m}, z, \gamma_L, \boldsymbol{\delta}, \eta) \\ &\qquad = p(\boldsymbol{c}_m|\boldsymbol{c}_{-m}, \gamma, \boldsymbol{\delta}) \times p(\boldsymbol{w}_m|\boldsymbol{c}, \boldsymbol{w}_{-m}, z, \eta) \end{aligned} \tag{5}$$

According to (5), the posterior probability of the path $\boldsymbol{c}_m$ is affected by two factor. The first factor is the prior on path $\boldsymbol{c}_m$ conditioned on all other paths. This prior can be divided into two situations. The first is selecting a table based on gCRP when there are tables occupied by local customers in the restaurant. In this case, gCRP ensures that the topic can expand horizontally. The second is selecting a table based on nCRP when there are no occupied tables, which enables vertical expansion of the topic hierarchy. The equation is organized as follows:

$$\begin{aligned} p(\boldsymbol{c}_m|\boldsymbol{c}_{-m}, \gamma, \boldsymbol{\delta}) &= p(\boldsymbol{c}^o_m|\boldsymbol{c}^o_{-m}, \boldsymbol{\delta}, \gamma) \times p(\boldsymbol{c}^e_m|\boldsymbol{c}^e_{-m}, \gamma) \\ &= \prod_{l=2}^{|\boldsymbol{c}^o_m|} p(c^o_{m,l} = k|\boldsymbol{c}^o_{-m}, \delta_l) \\ &\quad \times \prod_{l=|\boldsymbol{c}^o_m|+1}^{L} p(c^e_{m,l} = k|\boldsymbol{c}^e_{-m}, \gamma) \end{aligned} \tag{6}$$

The path $\boldsymbol{c}_m$ for document m consists of $\boldsymbol{c}^o_m$ and $\boldsymbol{c}^e_m$. $\boldsymbol{c}^o_m$ is a set of topics (can be labeled topics or generated latent topic) generated by gCRP from equation (2), when there are labeled topics in these levels. $\boldsymbol{c}^e_m$ is a set of latent topic generated by nCRP from equation (1), when there are no labeled topics in these levels.

Moveover, the second factor is the likelihood of obtaining the words for document m given a certain choice of path, which can be calculated as follows:

$$\begin{aligned} p(\boldsymbol{w}_m|\boldsymbol{c}, \boldsymbol{w}_{-m}, \boldsymbol{z}, \eta) &= \prod_{l=1}^{L} \frac{\Gamma(n^{(\cdot)}_{c_{m,l},-m} + V\eta)}{\prod_w \Gamma(n^{(w)}_{c_{m,l},-m} + \eta)} \\ &\quad \times \frac{\prod_w \Gamma(n^{(w)}_{c_{m,l},-m} + n^{(w)}_{c_{m,l},m} + \eta)}{\Gamma(n^{(\cdot)}_{c_{m,l},-m} + n^{(\cdot)}_{c_{m,l},m} + V\eta)} \end{aligned} \tag{7}$$

where $n^{(w)}_{c_{m,l},-m}$ represents the number of words w that have been allocated to the topic indexed by $c_{m,l}$, excluding those in the current document m, $\Gamma(\cdot)$ is the standard gamma function. From equation (6) and (7), the posterior probability of the path $\boldsymbol{c}_m$ is obtained, then the path for document m can be sampled.

2) TOPIC ALLOCATION SAMPLING

After obtaining the topic path of document m, we sample $z_{m,n}$, which is the topic allocation of the word n in document m. Given the current per-document path assignments of the whole corpus and the words in the document are observed, the conditional probability distribution can be represented as follows:

$$\begin{aligned} &p(z_{m,n} = k|\boldsymbol{z}_{-(m,n)}, \boldsymbol{w}, \boldsymbol{c}, \alpha, \eta) \\ &\qquad \propto p(z_{m,n} = k|\boldsymbol{z}_{-(m,n)}, \alpha) \\ &\qquad\quad \times p(w_{m,n} = v|\boldsymbol{z}, \boldsymbol{w}_{-(m,n)}, \boldsymbol{c}, \eta) \\ &\qquad = \frac{n^k_{-(m,n)} + \alpha}{n^{(\cdot)}_{-(m,n)} + |\boldsymbol{c}_m|\alpha} \cdot \frac{n^v_{k,-(m,n)} + \eta}{n^{(\cdot)}_{k,-(m,n)} + V\eta} \end{aligned} \tag{8}$$

where $n^k_{-(m,n)}$ is the number of words that have been allocated to topic k, excluding word n in document m, $n^v_{k,-(m,n)}$ is the number of times that token v has been generated under topic k, excluding word n in document m. According to equation (8), word n in document m can be allocated a topic in path $\boldsymbol{c}_m$.

3) GIBBS SAMPLING ALGORITHM

Having obtained all the conditional probability distribution, and giving the initial state of the Markov chain $\{\boldsymbol{c}^{(0)}_{1:M}, \boldsymbol{z}^{(0)}_{1:M}\}$, we iteratively sample each variable conditioned on the rest. After running the sampling process for sufficiently many iterations, the Markov chain approaches the stationary distribution. Algorithm 1 describes the parameter estimation process of HV-HTM model. Firstly, each document in the corpus samples a path, and each word in the document is also randomly allocated a topic. It is the initial state of the Gibbs sampling algorithm. Next, every document will resample a path and every word in the document will be reallocated a topic based on current state of the other document. Once all the documents have been sampled, an iteration of the Gibbs sampling is completed. In the convergence state, the topic-word distribution β_k and the document-topic distribution θ_m are estimated as follows:

$$\beta_{k,v} = \frac{n^v_k + \eta}{n^{(\cdot)}_k + V\eta} \tag{9}$$

$$\theta_{m,k} = \frac{n^k_m + \alpha}{n^{(\cdot)}_m + |\boldsymbol{c}_m|\alpha} \tag{10}$$

where n^v_k represents the number of times that token v appears in topic k, n^k_m represents the number of words in document m that have been allocated topic k.

V. EXPERIMENTS

In this section, we carry out various experiments on two real-world datasets, and demonstrate the effectiveness of the proposed model. Besides, our model is also compared with two state-of-the-art models on three evaluation metrics and two running performances. The experimental results are discussed and analyzed detailedly.

A. DATASETS

We download two very distinct datasets: one is originated from BBC News [29] and the other is from Yahoo! Answers (https://webscope.sandbox.yahoo.com/). BBC News dataset consists of 2,225 documents corresponding to stories in five topical areas (business, entertainment, politics, sport, technology) from 2004-2005. We refer the BBC News dataset as *BBC*. Yahoo! Answers dataset contains 142,627 questions and their answers. In addition to question and answer documents, the corpus contains the main-category, sub-category and category that are assigned to this question. We choose the question whose sub-category is different from its category. Then, we delete the categories and corresponding documents in main-category and sub-category if their number of documents is less 10. Finally, the Yahoo! Answers dataset contains 14,512 documents. We denote Yahoo! Answer dataset as *Y_Ans*.

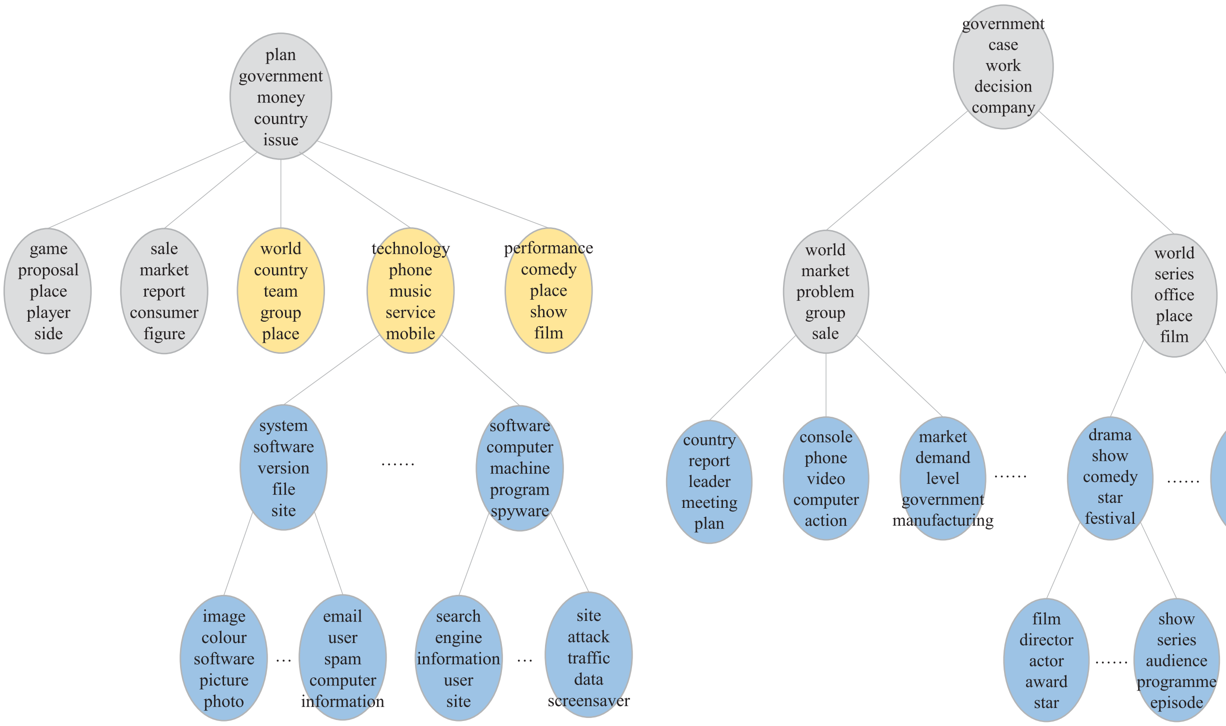
For both datasets, We remove stopwords with NLTK stopwords dictionary, delete words that are less than 3 letters, and filter out words other than nouns. Table 2 illustrates the statistics of the two datasets. According to the table, it can be seen that these two datasets present different characteristics: *Y_Ans* dataset has much more documents and labels than

Algorithm 1 The Parameter Estimation Process of HV-HTM Model

```
Input: Corpus − a collection of documents;
       L − the height of topical tree;
       Iter − the iteration number of Gibbs sampling;
       $\alpha, \eta, \gamma$ − the hyperparameters;
Output: TopicTree;
// Associate the distribution of vocabulary over topic $k$ with the node in TopicTree;
for each node $k \in$ TopicTree do
    draw a topic $\beta_k \sim Dir(\eta)$;
end for
// allocate initial state
for each document $m \in$ Corpus do
    Let $c_{m,1}$ be the root node;
    for each level $l \in \{2, \dots, L\}$ do
        if $c_{m,l-1}$ owns labeled child node then
            draw a node $c_{m,l}$ from gCRP with eq. (2);
        else
            draw a node $c_{m,l}$ from nCRP with eq. (1);
        end if
    end for
    obtain $c_m$
    draw an $L$-dim. topic proportion $\theta_m \sim Dir(\alpha)$;
    for each word $n \in \{1, \dots, N\}$ do
        draw a topic $z_{m,n}$ from $Mult(\theta_m)$;
        draw a word $w_{m,n}$ from $Mult(\beta_{c_{m,z_{m,n}}})$;
    end for
end for
// Gibbs sampling
for $i = 1$ to Iter do
    for each document $m \in$ Corpus do
        sample a new path $c_m$ with eq. (5);
        sample a new topic $z_{m,n}$ with eq. (8);
    end for
end for
return TopicTree;
```

TABLE 2. The statistics of the two datasets.

Dataset	*BBC*	*Y_Ans*
No. of labels	5	412
No. of paths	5	357
Max. height of the level	2	4
No. of documents	2,225	14,512
No. of tokens	7,226	32,052
Avg. number of documents per label	445	41
Avg. number of tokens per document	70	59


plan government money country issue
game proposal place player side
sale market report consumer figure
world country team group place
technology phone music service mobile
performance comedy place show film
system software version file site
......
software computer machine program spyware
image colour software picture photo
...
email user spam computer information
search engine information user site
...
site attack traffic data screensaver
government case work decision company
world market problem group sale
world series office place film
country report leader meeting plan
console phone video computer action
market demand level government manufacturing
......
drama show comedy star festival
......
lead title penalty player minute
film director actor award star
......
show series audience programme episode

(a) The topical tree discovered by HV-HTM

(b) The topical tree discovered by SSHLDA

FIGURE 5. **(a) A portion of topical hierarchy discovered by HV-HTM on *BBC* dataset. The whole tree contains 47 nodes. (b) A portion of topical hierarchy discovered by SSHLDA on *BBC* dataset. The whole tree contains 29 nodes. In both figure, the shaded nodes are topics with observed labels; the blue nodes are latent topics with vertical expansion; the yellow nodes are latent topics with horizontal expansion. The height of topical tree is 4. Each topic is represented by top five tokens.**

BBC dataset. The number of paths and maximal height of the level in *Y_Ans* dataset is also larger than *BBC* dataset, while *BBC* dataset has 10 times documents for each label and more tokens in each document than *Y_Ans* dataset.

All experiments are executed on our server with forty-eight Intel(R) Xeon(R) E5-2650 2.20GHz cores CPU, 386GB memory, and Ubuntu 16.04. For each model, the results are based on the states with a burn-in of 300 Gibbs sampling iterations. The optimal value of the three hyperparameters are determined by grid search. α is chosen from [5, 50] with step is 5. η is chosen from [0.01, 0.1] with step is 0.01, and γ is chosen from [0.1, 1] with step is 0.1. Finally, $\alpha = 10.0$, $\eta = 0.1$ and $\gamma = 1.0$ for *BBC* dataset. $\alpha = 15.0$, $\eta = 0.01$ and $\gamma = 1.0$ for *Y_Ans*.

B. CASE STUDY

In order to demonstrate the horizontal expansion capability of our proposed model, we merge directory *entertainment* and directory *sport*, and merge directories *business*, *politics* and *technology* of *BBC* dataset. Then two topical trees are constructed by our HV-HTM model and SSHLDA model respectively. Fig. 5 shows a portion of topical hierarchy discovered by our HV-HTM model and SSHLDA model.

Comparing Fig. 5 (a) and (b), we can summarize three major observations: (i) in level 2, SSHLDA only has two observed nodes, as SSHLDA directly selects an observed node from a multinomial distribution if nodes in this level have been observed. However, our HV-HTM generates five nodes, three of which are latent (as shown in yellow nodes). (ii) During constructing the hierarchical structure, HV-HTM relaxes the constraint of the observed nodes, it can generate more readable and comprehensible topics, while SSHLDA has a strong constraint when nodes in the level are observed, thus the topics corresponding to observed nodes are ambiguous. For example, in Fig. 5 (b), the two topics in level 2 are represented by ‘‘world market problem group sale’’ and ‘‘world series office place film’’ respectively. It is hard for us to relate them with two concepts. However, in Fig. 5 (a), it is easy to associate five nodes with five concepts ‘‘sports, business, country, technology, entertainment’’. (iii) When the height of the topical hierarchy of the two models are equal, HV-HTM can obtain fine-gained topics. For example, in Fig. 5 (a), the topics in level 4 are specific, such as ‘‘information retrieval’’ and ‘‘network attack’’. However, in Fig. 5 (b), the topics in level 4 are more general, such as ‘‘film’’ and ‘‘opera’’.

These significant observations further confirm that our proposed HV-HTM model can obtain better hierarchical structure than baseline methods.

C. EVALUATION METRICS AND EXPERIMENTAL RESULTS

To evaluate the effectiveness of our model, we employ perplexity, PMI and clustering ability as performance metrics and compare our model with two popular baseline models

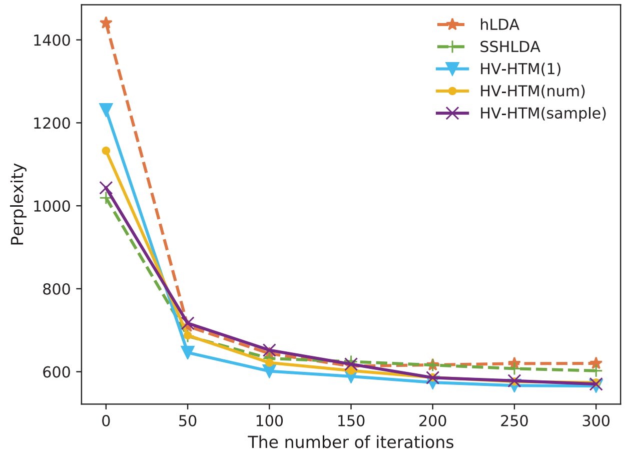


FIGURE 6. **The perplexities of hLDA, SSHLDA, HV-HTM(1), HV-HTM(num), and HV-HTM(sample) with different numbers of iterations. The results are run over the *BBC* dataset. The height of the topical tree is set to 4.**

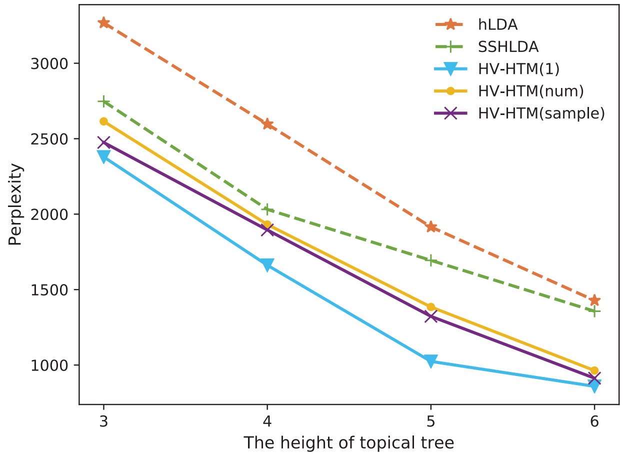


FIGURE 7. **The perplexities of hLDA, SSHLDA, HV-HTM(1), HV-HTM(num), and HV-HTM(sample) with different height of topical tree. The results are run over the *Y_Ans* dataset. The number of iterations is set to 300.**

quantitatively. Besides, we illustrate the runtime and memory usage of the three models.

1) EVALUATION ON PERPLEXITY

Perplexity is the most commonly used evaluation metric to evaluate the performance of topic model. Perplexity indicates the ability of topic model to generate unseen documents. Lower perplexity score means that topic model is more effective. For a collection of test documents, the perplexity score is calculated as follows:

$$Perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\log p(w_{dn})}{\sum_{d=1}^{M} N_d}\right\} \quad (11)$$

where N_d is the number of words in document d, and $p(w_{dn}) = \sum_z p(z|d)p(w_n|z)$ is the generation probability of word n in test document d. We randomly select 20% corpus as testing set and keep the remaining corpus as training set. HV-HTM model with three different strategies, described in Section 4, are conducted. The two state-of-the-art models, i.e. hLDA and SSHLDA, are also experimented to compare with HV-HTM.

Fig. 6 shows the perplexity scores of the three models over the *BBC* dataset with different numbers of iterations. As can be seen from the figure that the perplexities of three models eventually converge to steady values and our proposed model HV-HTM can achieve lower perplexity, compared to hLDA and SSHLDA. As SSHLDA and HV-HTM incorporate hierarchical label information into the topic modeling process, these two model have much lower initial perplexities, which means label information can provide valuable prior knowledge in the topic modeling process. It is also notable that although SSHLDA has the lowest initial perplexity value, the perplexity of SSHLDA shows a slower rate of decline. Its perplexity drops sightly after 100 iterations, while our proposed model HV-HTM shows a gradual downward tendency and reaches the lowest perplexity value.

Fig. 7 shows the perplexities scores of three models over the *Y_Ans* dataset with different height of the topical tree. We assume the maximal level of the observed labels are 3 and ignore the observed labels in level 4. In this Figure, same conclusion can be drawn that the perplexities of HV-HTM are lower than that of hLDA and SSHLDA at different height of the topical tree. Especially, when increasing the height of the topical tree, the performance improvement of HV-HTM is more obvious than that of SSHLDA, i.e. $L = \{5, 6\}$. Summarizing Fig. 6 and Fig.7, it means our proposed model HV-HTM can explore the relationship of the latent topics better than the state-of-the-art baseline models.

2) EVALUATION ON PMI

We also use PMI (Pointwise Mutual Information) score to evaluate whether the topics found by our model are reasonable and comprehensible. PMI describes the closeness of two words. If two words have a strong co-occurrence pattern, they may be highly correlated, and they have a higher PMI score. The PMI score of two words is defined as follows:

$$PMI(word_1, word_2) = \log\left(\frac{p(word_1, word_2)}{p(word_1)p(word_2)}\right) \quad (12)$$

where $p(word_1, word_2)$ is the co-occurrence frequency of $word_1$ and $word_2$ in the self-defined window, $p(word_i)$ is the frequency of word i in the corpus. We obtain the top 20 frequent words relevant to each topic, and calculate the PMI scores for each pair of words. The mean value of these PMI scores is set as the PMI score of topic k, as shown in equation (13).

$$PMI\text{–}topic_k = mean\{PMI(w_{ki}, w_{kj})\} \quad i, j \in [1, 20] \quad (13)$$

The PMI socre of the whole topical tree is set to the median value of the PMI scores of all topics in the topical tree, as shown in equation (14).

$$PMI\text{–}score_T = median\{PMI - topic_k\} \quad k \in [1, K] \quad (14)$$

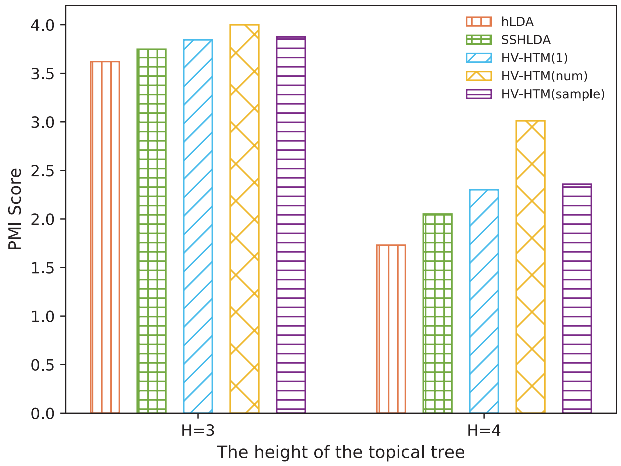


FIGURE 8. The comparison of PMI scores of hLDA, SSHLDA, HV-HTM(1), HV-HTM(num), and HV-HTM(sample) over different heights of the topical tree. The result is run over *BBC* dataset.

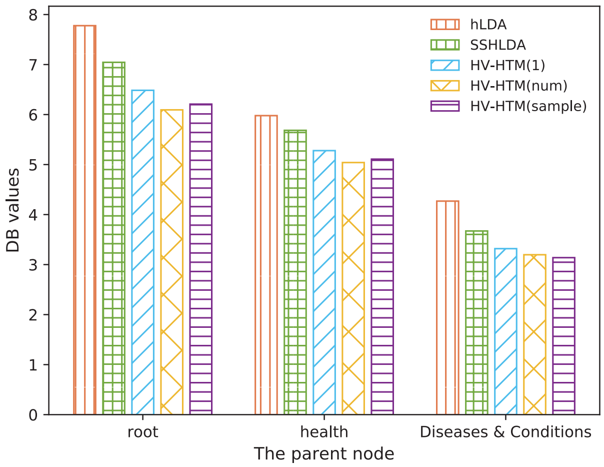


FIGURE 9. The comparison of DB values of hLDA, SSHLDA, HV-HTM(1), HV-HTM(num), and HV-HTM(sample) over different parent nodes at different level of the topical tree. The result is run over *Y_Ans* dataset.

We present the PMI scores of three models over the *BBC* dataset in Fig. 8. According to the figure, It can be found out that our model HV-HTM has higher PMI score, whether the height of the topical tree is 3 or 4, which means the words in each topic are highly relevant. In two different height of the topical tree, HV-HTM(num) has highest PMI score, and hLDA has lowest PMI score. When the height of the topical tree is 3, the PMI value of HV-HTM(num) reach 4.0, which is 10.5% and 7% higher than hLDA and SSHLDA. When the height of the topical tree is 4, the PMI value of HV-HTM(num) is 3.0, which achieve 73.4% and 52.3% improvement, compared to hLDA and SSHLDA. And with the increment of the height of the topical tree, the performance of hLDA and SSHLDA have a dramatic decrease, while the rate of decline of HV-HTM is relatively small, which indicates our model is more effective.

3) EVALUATION ON CLUSTERING ABILITY

Topic models have been successfully applied to clustering task. Each topic in the hierarchy and corresponding documents is a class. Thus we verify the reasonable of horizontal expansion of our model on clustering metric. A good clustering should create classes that instances in each class are closest to each other and the degree of separation between individual classes are larger. We employ Davies-Bouldin Index (DB) [30] to identify cluster overlap, the lower the DB value, the stronger the clustering capability of the model, which is defined as follows:

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\left(\frac{\overline{C_i}+\overline{C_j}}{||A_i-A_j||_2}\right) \tag{15}$$

where A_i is the centroid of class i, $||A_i - A_j||_2$ represents Euclidean distance between class i and j, and $\overline{C_i}$ is the mean distance of class i, which can be calculated as follows:

$$\overline{C_i} = \frac{1}{N_i}\sum_{n=1}^{N_i}||d_n - A_i||_2 \tag{16}$$

where d_n is the n*th* document in class i, and N_i is the number of documents in class i. In this experiment, we employ word2vec tool (which is published by Google in https://code.google.com/archive/p/word2vec/) to obtain continuous distributed representation of words [31], [32]. The project contains pre-trained vectors trained (300 dimensional vectors for 3 million words) on part of Google News dataset. The summation of the vector of the vector of each word in the document is used to represent the document. The height of the topical tree is 4. The children nodes of a parent node are treated as different classes. We choose root node in level 1, label "Health" in level 2 and label "Diseases & Conditions" in level 3 as parent nodes to demonstrate the clustering ability of different models. The DB values are illustrated in Fig. 9.

In Fig. 9, our HV-HTM model can achieve lower DB values than the baseline models over different parent nodes at different level. For example, when the parent node is root node, the DB value of hLDA is close to 8, and higher 22.1% than HV-HTM(num). The DB value of SSHLDA is above 7, which is higher 14.3% than HV-HTM. Similar results are showed when the parent is label "Health" or label "Disease & Conditions", which means our HV-HTM model can cluster documents into more appropriate classes.

We also count the number of child nodes for each parent node at different level and calculate their PMI scores, illustrated in Table 3. Numbers in bold font are the best results and numbers in underlined indicate the poorest results. As for root node, hLDA generates more than 60 child nodes, SSHLDA only has 12 child nodes, and our HV-HTM has about 40 child nodes. The number of child nodes of HV-HTM is more than that of SSHLDA and less than that of hLDA. It is because HV-HTM relaxes the restrict that the latent

TABLE 3. The PMI scores over different parent nodes on *Y_Ans* dataset.

Model	root	Health	Disease & Conditions
hLDA	3.505	3.058	2.961
SSHLDA	3.818	3.645	3.243
HV-HTM(1)	3.961	3.762	3.620
HV-HTM(num)	**4.214**	3.877	3.674
HV-HTM(sample)	4.109	**3.899**	**3.802**

TABLE 4. The runtime and memory usage of three models on *BBC* dataset.

Model	runtime(min)		memory usage(MB)	
	3[1]	4	3	4
hLDA	95	118	14.87	26.55
SSHLDA	83	93	13.02	21.01
HV-HTM	92	105	14.27	23.16

[1] The number here represents the height of the topical tree.

topics need to choose from one of observed labels. Therefore, HV-HTM can generate more than 12 child nodes. Compared to hLDA, HV-HTM takes the observed label information into modeling process, which controls the generation of new child nodes. Therefore, the number of child nodes of HV-HTM is less than hLDA. Moveover, the PMI scores are 3.505, 3.818, 3.961, 4.214 and 4.109 respectively. Our HV-HTM model has significant improvement. When parent nodes are ''Health'' and ''Disease & Conditions'', there are similar results. These results further prove that the topic horizontal expansion based on observed label information is proper.

4) EVALUATION ON RUNTIME AND MEMORY USAGE

In addition to evaluating our model on three evaluation metrics, the runtime and memory usage of three models are compared in this section. The runtime and memory usage are mainly affected by the size of the topical tree, and larger topical tree means larger memory usage and longer processing time. Table 4 and Table 5 present the results on *BBC* dataset and *Y_Ans* dataset respectively.

On both dataset, SSHLDA has the lowest runtime and memory usage, and hLDA has the highest runtime and memory usage, with HV-HTM in the middle. For the reason that hLDA do not take advantage of observed label information that hLDA generates a large topical tree, while, the observed label information restricts SSHLDA from generating latent topic that SSHLDA generates fewer topics. HV-HTM not only incorporates observed label information into topic generation process, but also maintains the flexibility of topic expansion. Therefore, HV-HTM generates a medium-size topical tree. Moveover, in Table 5, as the height of the topical tree increases, the gap between the runtime and memory usage of HV-HTM and SSHLDA gradually narrow. However, the runtime and memory usage of hLDA show a dramatic increase, much more than that of SSHLDA and HV-HTM. These results indicate that HV-HTM has the same running performance as SSHLDA and is much better than hLDA.

TABLE 5. The runtime and memory usage of three models on *Y_Ans* dataset.

Model	runtime(min)			memory usage(MB)		
	4[1]	5	6	4	5	6
hLDA	215	285	390	59.39	75.77	108.54
SSHLDA	173	222	357	51.02	64.51	88.06
HV-HTM	185	236	362	56.32	70.68	93.18

[1] The number here represents the height of the topical tree.

VI. CONCLUSION

In this paper, we focus on hierarchical topic models incorporating the observed hierarchical label information and how to expand the topical tree horizontally and vertically. We presented the novel horizontal and vertical hierarchical topic model, called HV-HTM. This model first select a labeled node or generate a new node in an observed hierarchy through a general version of the Chinese restaurant process, which incorporates label information into the topic generation process and keep the flexibility of the horizontal topic expansion. Then this model generates a sub-hierarchy in deeper vertical level. We conducted experiments on BBC news and Yahoo! Answers datasets, and evaluated the performance of HV-HTM in terms of Perplexity, PMI scores and DB values. The experimental results showed that HV-HTM has a significant improvement on predictive ability, compared to the state-of-the-art models, and it can also obtain more interpretable hierarchical structure.

In the future, we will apply the proposed models to other real-world corpus, such as MicroBlog and Twitter data. We will also continue to explore novel topic models for other kinds of label information or prior knowledge.

REFERENCES

[1] H. Jelodar, "Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019.

[2] L. Liu, "An overview of hierarchical topic modeling," in *Proc. 8th Int. Conf Intell. Hum.-Mach. Syst. Cybern.*, Hangzhou, China, Aug. 2016, pp. 391–394.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[4] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertain. Artif. Intell.*, San Francisco, CA, USA, 1999, pp. 289–296.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[6] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, Aug. 2006, pp. 424–433.

[7] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 113–120.

[8] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2005, pp. 147–154.

[9] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertain. Artif. Intell.*, Arlington, VA, USA, 2004, pp. 487–494.

[10] W. X. Zhao, "Comparing Twitter and traditional media using topic models," in *Proc. Eur. Conf. Inf. Retr.*, Dublin, Ireland, Apr. 2011, pp. 338–349.

[11] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in *Proc. Adv. Neural Inf. Process. Syst.*, Whistler, BC, Canada, Dec. 2003, pp. 17–24.

[12] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, Jan. 2010, Art. no. 7.

[13] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 577–584.

[14] D. M. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with Pachinko allocation," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 633–640.

[15] S. Kim, J. Zhang, Z. Chen, A. H. Oh, and S. Liu, "A hierarchical aspect-sentiment model for Online reviews," in *Proc. 27th AAAI Conf. Artif. Intell.*, Bellevue, WA, USA, Jul. 2013, pp. 526–533.

[16] L. Zhu, Y. He, and D. Zhou, "Hierarchical viewpoint discovery from Tweets using Bayesian modelling," *Expert Syst. Appl.* vol. 116, pp. 430–438, Feb. 2019.

[17] S.-J. Shin and I.-C. Moon, "Guided HTM: Hierarchical topic model with Dirichlet forest priors," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 2, pp. 330–343, Feb. 2017.

[18] D. Yu, D. Xu, D. Wang, and Z. Ni, "Hierarchical topic modeling of Twitter data for online analytical processing," *IEEE Access*, vol. 7, pp. 12373–12385, 2019.

[19] Y. Petinot, K. R. McKeown, and K. Thadani, "A hierarchical model of Web summaries," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, Portland, OR, USA, Jun. 2011, pp. 670–675.

[20] A. Ahmed, L. Hong, and A. Smola, "Nested chinese restaurant franchise process: Applications to user tracking and document modeling," in *Proc. 30th Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. 1426–1434.

[21] K. W. Lim, W. L. Buntine, C. Chen, and L. Du, "Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes," *Int. J. Approx. Reasoning*, vol. 78, pp. 172–191, Nov. 2016.

[22] A. J. Perotte, F. D. Wood, N. Elhadad, and N. Bartlett, "Hierarchically supervised latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 2609–2617.

[23] V. Nguyen, J. L. Boyd-Graber, and P. Resnik, "Lexical and hierarchical topic regression," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 1106–1114.

[24] X. Mao, Z. Ming, T. Chua, S. Li, H. Yan, and X. Li, "SSHLDA: A semi-supervised hierarchical topic model," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, Jeju Island, South Korea, Jul. 2012, pp. 800–809.

[25] W. Wang, H. Xu, W. Yang, and X. Huang, "Constrained-hLDA for topic discovery in Chinese microblogs," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Tainan, Taiwan, May 2014, pp. 608–619.

[26] K. Yamamoto, K. Eguchi, and A. Takasu, "Hierarchical topic models for expanding category hierarchies," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, Kyoto, Japan, Mar. 2019, pp. 1–8.

[27] D. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII*, Berlin, Germany: Springer, 1985, pp. 1–198.

[28] X. Zhou, J. Ouyang, and X. Li, "A more time-efficient gibbs sampling algorithm based on SparseLDA for latent Dirichlet allocation," *Intell. Data Anal.*, vol. 22, no. 6, pp. 1227–1257, 2018.

[29] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. 30th Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 377–384.

[30] A. Fahad, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

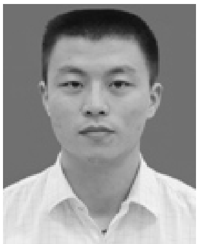
[32] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Nerual Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.

XI ZOU was born in Jingzhou, Hubei, China. He received the B.S. degree in computer science and technology from Hohai University, Nanjing, China, in 2017, where he is currently pursuing the M.S. degree in computer science and technology. He has participated in the National Key Research and Development Program of China and several government-funded projects. His research interests include machine learning, data mining, natural language processing, and knowledge graph construction.

YUELONG ZHU was born in Yancheng, Jiangsu, China. He is currently a Professor with the School of Computer and Information, Hohai University. His main research interests include intelligent information processing, data mining, and water resources informatization.

JUN FENG was born in Xuzhou, Jiangsu, China, in 1969. She received the B.S. and M.S. degrees in computer science and technology from Hohai University, China, in 1991 and 1994, respectively, and the Ph.D. degree in information engineering from the University of Nagoya, Japan, in 2004. She is currently a Professor with the School of Computer and Information, Hohai University. She has authored the book *Index and Query Methods in Road Networks* (Springer, 2015). Her research interests include data management, spatiotemporal indexing and search methods, ITS, and domain data mining.

JIAMIN LU was born in Nantong, Jiangsu, China, in 1983. He received the B.S. and M.S. degrees in computer science and technology from Hohai University, Nanjing, China, in 2004 and 2008, respectively, and the Ph.D. degree in information science from FernUniversität, Hagen, Germany, in 2014. He is currently a Lecturer with the School of Computer and Information, Hohai University. His research interests include parallel processing on MOD, cloud infrastructure, and knowledge graph construction.



XIAODONG LI was born in Nanjing, Jiangsu, China. He received the B.S. degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, in 2014. He is currently an Associate Professor with the School of Computer and Information, Hohai University. He has served as a Principal Investigator of a sub-project in the National Key Research and Development Program of China and a project in the National Natural Science Foundation of China. His research interests include artificial intelligence, machine learning, information retrieval, and data mining, especially big data applications to algorithmic trading in finance and water resource management in hydrology.