IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 9, 2020, accepted March 19, 2020, date of publication March 25, 2020, date of current version April 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983265

A Top-Down Binary Hierarchical Topic Model for Biomedical Literature

XIAOGUANG LIN[1,2,3], MINGXUAN LIU[2,3], AND JU ZHANG[2,3]
[1]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China
[2]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Ju Zhang (zhangju@cigit.ac.cn)

ABSTRACT Over the past two decades, a number of advances in topic modeling have produced sophisticated models that are capable of generating topic hierarchies. In particular, hierarchical Latent Dirichlet Allocation (hLDA) builds a topic tree based on the nested Chinese Restaurant Process (nCRP) or other sampling processes to generate a topic hierarchy that allows arbitrarily large branch structures and adaptive dataset growth. In addition, hierarchical topic models based on the latent tree model, such as Hierarchical Latent Tree Analysis (HLTA), have been developed over the last five years. However, these models do not work well in cases with millions of documents and hundreds of thousands of terms. In addition, the topic trees generated by these models are always poorly interpretable, and the relationships among topics in different levels are relatively simple. The biomedical literature, including Medline abstracts, has large-scale documents in two major categories: biological laboratory research and medical clinical research. We propose a top-down binary hierarchical topic model (biHTM) for biomedical literature by iteratively applying a flat topic model and adaptively processing subtrees of the hierarchy. The biHTM topic hierarchy of complete Medline abstracts with more than 14 topic node levels shows good bimodality and interpretability. Compared to hLDA and HLTA, biHTM shows promising results in experiments assessed in terms of runtime and quality.

INDEX TERMS Topic model, topic hierarchy, binary modality, biomedical literature, text mining.

I. INTRODUCTION

In the last two decades, the biomedical literature has grown exponentially, which has created an enormous challenge for life science researchers and healthcare professionals attempting to stay up to date with developments in their field [1]. These vast collections of publications offer an excellent opportunity for text mining, i.e., the automatic discovery of knowledge and deep semantic retrieval. Some exceptional natural language processing (NLP) methods, such as topic models, can be used to text mine the biomedical literature.

Topic models are generally used to preprocess a document collection, and the topics and per-document topic allocations are fed to downstream applications such as document classification methods, novel word sense detection methods and machine translation methods [2]. The predominant topic model is Latent Dirichlet Allocation (LDA) [3], which is an unsupervised three-layer Bayesian probability model that highlights the relationships among terms, documents, and potential semantic topics.

Most flat topic models, such as LDA, treat document subjects as a set of probability distributions with no direct relationships between topics. These models can be used to mine the topics in the corpus; however, the associations and hierarchies between topics cannot be found. Several hierarchical topic models have been proposed to obtain the topic hierarchy, including hierarchical Latent Dirichlet Allocation (hLDA) and Hierarchical Latent Tree Analysis (HLTA). These models build a topic hierarchy that allows for arbitrarily large branch structures and adaptive dataset growth. In addition, they have been successfully applied for document modeling, online advertising and microblog location prediction and outperformed flat topic models [4].

The hLDA and its variants have two main implementations: the nested Chinese Restaurant Process (nCRP) [5], [6] and the nested Hierarchical Dirichlet Process (nHDP) [7]. The hLDA and its variants [6], [8]–[10] cluster documents, establish a hierarchical structure, facilitate the learning of topic hierarchy information, and mine potential topic

The associate editor coordinating the review of this manuscript and approving it for publication was Yu Zhang.

information. These models need to maintain a full topic hierarchy in each sampling iteration, and it usually takes at least 100 iterations to yield a stable hLDA topic tree. Therefore, it is difficult for hLDA and its variants to process massive text corpora. In addition, the hLDA topic results of large-scale corpora are usually ambiguous and enormous and contain redundant information.

HLTA and other latent tree models treat words as binary variables, and each word is allowed to appear in only one branch of a topic hierarchy. However, there are many ambiguous words, and a single word, when combined with different words, can result in different topics. For example, ''*coronavirus*, *patients*, *treatment*'' may be related to clinical treatment of coronavirus, while ''*coronavirus*, *bat*, *effect*'' may be related biological research of the coronavirus. Thus, topic hierarchies generated by these latent tree models may be not as comprehensive as those generate by hLDA. Similar to hLDA and its variants, HLTA and other latent tree models are probabilistic generative models, and they all need to involve latent variables. The number of levels in a topic hierarchy resulting from HLTA may depend on the number of words in the vocabulary. In fact, a topic tree with more than six levels resulting from these models is not common.

The biomedical literature has large documents, such as the documents from Medline,[1] a preeminent bibliographic database that contains more than 22 million article abstracts from life sciences journals and adds approximately 2000–4000 abstracts every day. Existing hierarchical topic models cannot mine such documents well. The biomedical literature includes two major categories: biological laboratory research and medical clinical research. Thus, a binary hierarchical topic model may be more suitable for biomedical literature than other models. In addition, the levels of a binary topic hierarchy can be much deeper than those of topic hierarchies with more than two topics in each level. The path from root node to leaf nodes in a binary tree model can be much longer than other multi-tree models. Therefore, the relationships among different levels of topics in a binary topic hierarchy will be fairly rich.

We propose a top-down binary hierarchical topic model (biHTM) for the biomedical literature by iteratively applying a flat topic model (such as LDA), and adaptively processing the subtrees of the hierarchy. This method is a heuristic generative method, different from other probabilistic generative methods, and it could quickly generate the topic hierarchy from top to bottom without using latent variables. This method is adaptive with few hyper-parameters. To avoid interference between different contents and mine local detail information, we split the corpus into a set of doculets by using a 3-sentence sliding window and then recursively learn an adaptive top-down topic hierarchy.

This paper is organized as follows. Related work is introduced in section II. We present our method in detail in section III. The experiments and results of biHTM trained with complete Medline abstracts are shown in section IV. Section V compares biHTM, hLDA, and HLTA in terms of topic interpretability, topic quality and efficiency. Finally, section VI concludes this paper.

[1]https://www.ncbi.nlm.nih.gov/pubmed

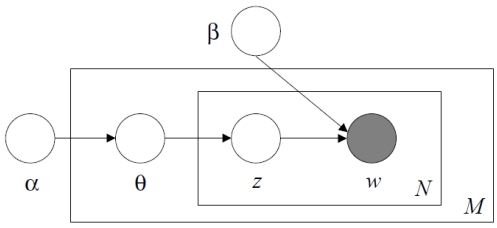


FIGURE 1. Graphical model representation of LDA.

II. RELATED WORK

Topic detection has been one of the most active research areas in machine learning over the past few decades, and the main topic models include latent semantic analysis (LSA) [11], probabilistic LSA (pLSA) [12], and LDA [3]. In a probabilistic flat topic model, each document in a collection of D documents is modeled as a multinomial distribution over T topics, where each topic is a multinomial distribution over W words. Typically, only a small number of words are important (i.e., have a high likelihood) in each topic, and only a small number of topics are present in each document.

The most commonly used topic model is LDA [3], [13], [14], which is represented as a probabilistic graphical model in Figure 1, and it has shown great success in various NLP tasks for discovering latent topics in the biomedical literature [15], [16]. However, topics are naturally organized in a hierarchy [4], and the above models are flat topic models that cannot capture the hierarchical information of the topics.

To address the above problem, some researchers have tried to extend the traditional topic modeling techniques to obtain the hierarchical information of the topics. Blei *et al.* proposed the hLDA model [5], [6], which can automatically learn the hierarchical topic structure with Gibbs sampling by utilizing a nested Chinese Restaurant Process (nCRP) prior in an unsupervised way. Mimno *et al.* proposed the hierarchical Pachinko allocation model (HPAM) [17] based on Pachinko allocation to generate the topic hierarchy. In HPAM, a document is generated from a distribution over the topics at the leaves (the lowest level of nodes) of a topic hierarchy. Paisley *et al.* proposed the nested Hierarchical Dirichlet Process (nHDP) instead of nCRP in hLDA [7]. The nHDP allows each word to follow its own path to a topic node according to a per-document distribution over the paths on a shared tree.

Dai *et al.* proposed the supervised hierarchical Dirichlet process (sHDP), a nonparametric generative model for the joint distribution of a group of observations and a response variable directly associated with that whole group [18]. Based on hLDA, Mao *et al.* proposed Semi-Supervised Hierarchical Latent Dirichlet Allocation (SSHLDA), which assumed that some topics in different levels in a topic hierarchy are known before modeling and used such known topics as

prior knowledge to generate the corresponding structure in a topic hierarchy [9]. Wang *et al.* also proposed a semi-supervised hierarchical topic model that aims to explore reasonable topics in the data space by incorporating constraints that are extracted automatically in the modeling process [19]. Xu *et al.* proposed a novel knowledge-based hierarchical topic model (KHTM), which can incorporate prior knowledge into topic hierarchy building to solve the problem of weak topic hierarchies [20]. Chen *et al.* proposed a partially collapsed Gibbs sampling algorithm and a supercomputer to apply nHDP [7], [21] on some large complete datasets and produced some very interesting topic hierarchies [4]. Zou *et al.* proposed a novel hierarchical topic model (HV-HTM) to incorporate the observed hierarchical label information into the topic generation process while maintaining the flexibility of the horizontal and vertical expansion of the hierarchical structure in the modeling process [22]. Yu *et al.* proposed a topic model called twitter hierarchical latent Dirichlet allocation (thLDA) to automatically mine the hierarchical dimension of tweets' topics, which can be further employed to text OLAP on tweets [23].

The above LDA-based hierarchical topic models can obtain relatively comprehensive hierarchical topic information. However, training these models can take substantial amounts of time, and they require the user to provide the structure of a hierarchy, including the number of latent levels and the number of nodes at each level. The number of latent levels is usually set to 3 based on efficiency considerations.

Chen *et al.* described a novel hierarchical topic detection method called HLTA [8], [24], which explores a topic hierarchy of word co-occurrence relationships. Another latent tree method, the Correlation-Explanation method, by Steeg *et al.* uses information theory arguments [25], [26]. The method is named correlation explanation (CorEx), and the latent variables are used to maximally explain the correlations in the layer below, thus forming optimally informative hierarchical topics. These hierarchical topic models based on the latent tree method can obtain hierarchical topic information efficiently; however, they extract a specific number of key words as the observation variables, and each word appears in only one branch of the topic hierarchy. Thus, topic hierarchies generated by these latent tree models may be not as comprehensive as those generated by hLDA and its variants.

Most of the above-mentioned hierarchical topic models are probabilistic generative models. They all involve latent variables and construct the topic hierarchy using different processes. In general, the number of levels of these topic hierarchies will be less than 5, and the relationships among topics in different levels are relatively simple.

In recent years, hierarchical topic models have been successfully applied to microblog location prediction [27], web hierarchical topic detection [28], the identification of correlations between topics and social networks [29], knowledge mining [20] and social comment short text analytics [23].

With the rapid growth of the biomedical literature, it is desirable to harvest information and knowledge in the literature via text mining [30]–[32]. There is minimal research on hierarchical topic models for medical literature.

In this paper, to efficiently mine large-scale medical literature corpora, we proposed an adaptive top-down binary hierarchical topic model, called biHTM, and the resulting biHTM topic tree of Medline abstracts supported our conjecture about the binary characteristics of the biomedical literature.

III. METHOD

A. DOCULET

A biomedical document consists of many paragraphs, and each paragraph consists of several sentences. In equations (1) and (2), O denotes a document, P_i denotes a paragraph, and S_{ij} denotes a sentence. n is the number of paragraphs in the document, and m is the number of sentences in a paragraph.

$$\mathrm{O} = \{P_1, \cdots, P_i, \cdots, P_n\}, \quad 1 \le i \le n \tag{1}$$

$$P_i = \{S_{i1}, \cdots, S_{ij}, \cdots, S_{im}\}, \quad 1 \le j \le m \tag{2}$$

Definition: A **doculet** d is a sentence set containing three consecutive sentences in a paragraph.

$$d_{ij} = \{S_{ij}, S_{i(j+1)}, S_{i(j+2)}\} \tag{3}$$

$$D_i = \{d_{i1}, \cdots, d_{ij}, \cdots, d_{i(m-2)}\} \tag{4}$$

$$D = \{D_1, \cdots, D_i, \cdots, D_n\} \tag{5}$$

d_{ij} denotes one of the doculets in a paragraph, D_i denotes all the doculets in a paragraph, and D denotes all the doculet sets in an article. If a paragraph has one or two sentences, then these sentences should be integrated in the next paragraph.

Some studies attempt to apply LDA to sentences instead of documents to improve topic quality [33], [34]. In most cases, one or two sentences have less useful information to analyze, and four or more sentences probably have a mixture of different types of information. In this paper, the biomedical literature corpus is broken into a new input corpus of doculets for training by using a 3-sentence sliding window. Compared to natural paragraphs, doculets can reduce the ambiguity of topics and improve the quality of topic hierarchy generation.

B. TOPIC PROPORTION

A corpus of doculets C is trained with LDA to obtain 2 topics T, where each topic t contains the most likely dictionary words h from the doculet corpus C after LDA estimation. The relations among the document, topics, and words are described mathematically in reference [3]

$$T = \{t_1, t_2\} \tag{6}$$

$$t = \{W_1, W_2, \cdots, W_h\} \tag{7}$$

Each word W in a doculet d can be categorized as one of two topics: t_1 or t_2.

Definition: The **topic proportion** x of a doculet is the number of dictionary words categorized as t_1 divided by the number of total dictionary words in the doculet.

$$x = \frac{\mathrm{num}\left(w'_1, \cdots, w'_{z'}\right)}{\mathrm{num}(w_1, \cdots, w_z)}, \left\{w'_1, \cdots, w'_{z'}\right\} \subseteq \{w_1, \cdots, w_z\} \tag{8}$$

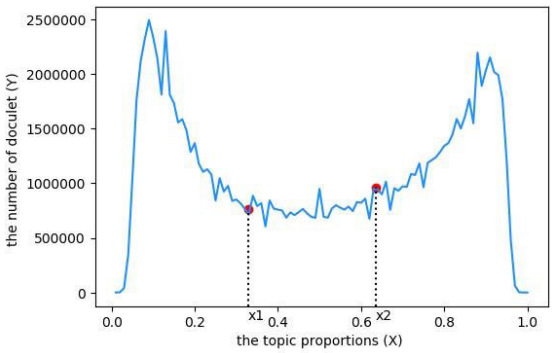


FIGURE 2. The distribution of doculets for the complete Medline abstracts.

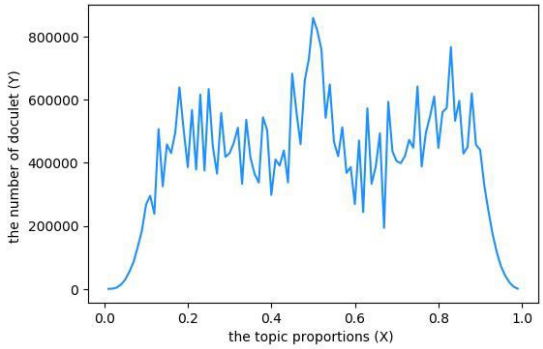


FIGURE 3. The distribution of doculets for the Yelp dataset.

In equation (8), w' denotes the dictionary words in the doculet categorized as t_1, and w denotes the dictionary words in the doculet categorized and uncategorized as t_1. For each doculet, x denotes the proportion of words for topic t_1, and 1-x denotes the proportion of words for topic t_2. If x is closer to 0, then the doculet is more similar to topic t_2. In contrast, a doculet is more similar to topic t_1 if x is closer to 1. Obviously, x is a decimal value between 0 and 1. In this paper, x can take only 101 values: those uniformly distributed between 0 and 1; that is

$$x \in \{0, 0.01, 0.02, \ldots, 0.99, 1\} \tag{9}$$

All the values computed with equation (8) should be rounded to the values in equation (9).

Figure 2 shows the distribution of doculets for the corpus of complete Medline abstracts. The values on the x axis are topic proportions, and the values on the y axis are the number of doculets. Figure 3 shows the distribution of the doculets for the Yelp comment corpus.[2] Apparently, if relatively high numbers of doculets are distributed near the ends of the x axis, this distribution shows bimodality. Figure 2 shows good bimodality, while Figure 3 shows no bimodality.

C. biHTM

The biHTM is a top-down hierarchical topic model recursively trained with a corpus of doculets to obtain two topics

[2] https://www.yelp.com/dataset

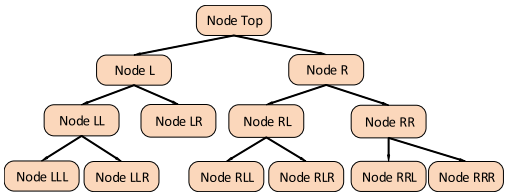


FIGURE 4. A biHTM topic tree example.

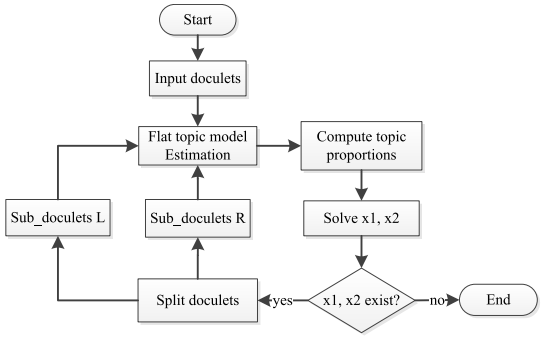


FIGURE 5. The biHTM topic tree learning process.

with LDA. The doculets are separately divided into two sub-nodes according to the topic proportions after each LDA estimation. An example of a 4-level biHTM topic tree is shown in Figure 4.

A topic tree node can be defined as a tuple:

$$N = \langle D, T, N_l, N_r \rangle \tag{10}$$

$$N_l = \langle D_l, T_l, N_{ll}, N_{lr} \rangle \tag{11}$$

$$N_r = \langle D_r, T_r, N_{rl}, N_{rr} \rangle \tag{12}$$

We treat N as the root node, where D denotes a corpus of doculets; T denotes the two topics generated with LDA, and N_l and N_r are the sub-nodes of N.

The doculets with low topic proportions (close to 0) in D are assigned to N_r, and the doculets with high topic proportions (close to 1) in D are assigned to N_l. However, the doculets with moderate topic proportions (close to 0.5) in D should be abandoned because these doculets have a large amount of mixed information and are difficult to categorize into one topic.

D. LEARNING PROCESS

Figure 5 shows the learning process of the biHTM topic tree. Two topics are generated in each LDA estimation step with the current doculet sets. Then, the topic proportions of each doculet can be computed with the results of LDA estimation. We find two appropriate points in the distribution of doculets, such as x_1 and x_2, as shown in Figure 2. If x_1 and x_2 exist, then the doculets distributed in $[x_1, x_2]$ are abandoned, the doculets in $[0, x_1)$ are classified as sub-doculets R, and the doculets in $(x_2, 1]$ are classified as sub-doculets L. Then, we separately train these two sub-doculets with LDA. We repeat the above procedure to obtain the final tree.

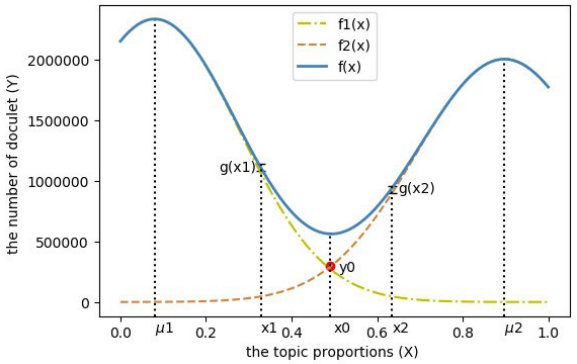


FIGURE 6. Mixture Gaussian curve fitting the distribution of doculets in each node.

It is crucial to determine x_1 and x_2; the following paragraph describes the method used to determine them.

We use $F(x)$ to denote the distribution of doculets over the interval [0, 1]; in other words, $F(x)$ denotes the number of doculets on the topic proportion x. From section B, it can be seen that $F(x)$ is discrete with 101 points. An example of $F(x)$ is shown in Figure 2. There are two important parameters in each node: x_1 and x_2. The doculets for which the topic proportion is less than x_1 are assigned to N_l, and the doculets for which the topic proportion is greater than x_2 are assigned to N_r. If x_1 and x_2 are unsuitable or unavailable, then node N will not continue to expand and will become a leaf node.

The curve shown in Figure 2 has two peaks and thus shows good bimodality. We use a Gaussian mixture curve $f(x)$ with two Gaussian distributions $f_1(x)$ and $f_2(x)$ to fit $F(x)$, as shown in Figure 6. In this paper, we fit a 2-peak Gaussian mixture curve using nonlinear iterative curve fitting [35].

$$\begin{aligned} f(x) &= f_1(x) + f_2(x) = \alpha N\left(\mu_1, \sigma_1^2\right) + \beta N\left(\mu_2, \sigma_2^2\right) \\ &= \frac{\alpha}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \\ &\quad + \frac{\beta}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right), \; x \in (0,1). \end{aligned} \tag{13}$$

We set the point (x_0, y_0) as the intersection of curves f_1 and f_2; that is

$$y_0 = f_1(x_0) = f_2(x_0). \tag{14}$$

From equations (13) and (14), we get

$$\begin{aligned} y_0 &= \frac{\alpha}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_0-\mu_1)^2}{2\sigma_1^2}\right) \\ &= \frac{\beta}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_0-\mu_2)^2}{2\sigma_2^2}\right). \end{aligned} \tag{15}$$

Then, x_0 can be calculated based on equation (15), (16), as shown at the bottom of this page.

We want to find the two values x_1 and x_2 on the axis x; by setting $\mu_1 < x_1 < x_0 < x_2 < \mu_2$, as shown in Figure 6, the doculets distributed over the ranges $[0, x_1)$ and $(x_2, 1]$ will be divided and transferred to the next levels to separately estimate flat topics, and the doculets distributed over the range $[x_1, x_2]$ will be abandoned. We solve x_1 and x_2 as follows.

We set

$$g(x) = \begin{cases} f(x) - f_1(x) = f_2(x), & x \le x_0 \\ f(x) - f_2(x) = f_1(x), & x \ge x_0 \end{cases} \tag{17}$$

$$g(x) \in (0, y_0], \quad x \in [0, 1] \tag{18}$$

$$g(x_1) = g(x_2), \quad x_2 - x_1 \in (0, 1]. \tag{19}$$

To abandon as few doculets as possible and to abandon doculets that are difficult to any categorize to the greatest extend possible, we obtain a multiobjective problem for determining x_1 and x_2:

$$\begin{cases} \arg\min\limits_{x_1, x_2} \{g(x_1) = g(x_2) \in (0, y_0]\} \\ \arg\min\limits_{x_1, x_2} \{x_2 - x_1 \in (0, 1]\}. \end{cases} \tag{20}$$

According to equation (20), $x_2 - x_1$ should be as small as possible and $g(x_1)$ or $g(x_2)$ should be as small as possible.

We suppose that the weight of $g(x_1)$ and $g(x_2)$ is equal to the weight of $x_2 - x_1$; thus, equation (20) can be transformed into equation (21) with a normalization factor.

$$\arg\min_{x_1, x_2} \left(v = 0.5\frac{g(x_1)}{y_0} + 0.5(x_2 - x_1) \right). \tag{21}$$

Unfortunately, equation (21) has no analytic solution. However, we can find the most appropriate numerical solution by searching a number of sample values. The searching algorithm is shown as Algorithm 1.

The termination conditions for the current recursion include: the fitting curve is not found; appropriate x_1 and x_2 are not determined; and $F(x_0) \ge \min(f_1(\mu_1), f_2(\mu_2))$.

IV. EXPERIMENT

A. MEDLINE ABSTRACT CORPUS

Medline is a bibliographic database of life science and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and healthcare. Medline abstracts contain more than 25 million abstracts from 5639 selected publications covering biomedicine and health from 1950 to the present. Each Medline abstract file is provided in a normalized.xml

$$x_0 = \frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2 \pm \sqrt{\left(\mu_1\sigma_2^2 - \mu_2\sigma_1^2\right)^2 - \left(\sigma_1^2 - \sigma_2^2\right)\left(\mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2 - 2\sigma_1^2\sigma_2^2\left(\ln\beta\sigma_1 - \ln\alpha\sigma_2\right)\right)}}{\sigma_1^2 - \sigma_2^2}. \tag{16}$$

TABLE 1. Key parameters of the first 3-level biHTM tree nodes for the complete Medline abstracts.

Level	μ_1	μ_2	σ_1	σ_2	α	β	$g(x_1)$	(x_1, x_2)
Top	0.090	0.900	0.185	0.190	920298	858737	57692	(0.40, 0.58)
L	0.100	0.890	0.200	0.187	297486	307098	10984	(0.36, 0.67)
R	0.120	0.881	0.190	0.189	249123	290643	12662	(0.35, 0.64)
LL	0.090	0.886	0.189	0.192	86700	120137	3936	(0.32, 0.61)
LR	0.130	0.923	0.187	0.163	90302	100302	2527	(0.43, 0.68)
RL	0.129	0.890	0.221	0.252	83364	124000	11088	(0.29, 0.64)
RR	0.110	0.904	0.188	0.191	116056	75095	3489	(0.37, 0.66)

Algorithm 1 The Search Process of x_1 and x_2

Input: $X_{0\sim9999}-10000$ values uniformly distributed along the axis x
α, μ_1, σ_1- Gaussian curve f_1
β, μ_2, σ_2- Gaussian curve f_2
Output: x_1, x_2

```
Define a variable x1, x2, v', and initialize x1, x2, v' to 1;
Calculate x0, y0;   //equation (15) and (16)
for m: = (μ1, x0) ∈ X do
  for n: = (x0, μ2) ∈ X do
    compute g(m);   //equation (13) and (17)
    compute g(n);   //equation (13) and (17)
    if (1 - min(g(m), g(n)) / max(g(m), g(n)) ) < t then
    //t is a threshold, we set t = 0.01 in this paper
      compute v;   //equation (21)
      if v < v' then
        v' := v;
        x1 := m;
        x2 := n;
      end if
    end if
  end for
end for
if v' != 1 then
  return x1, x2;
end if
```

format and contains a large amount of useful information, such as journal metadata, article metadata, title, keyword, abstract text, and reference information.

B. MALLET

Mallet is an open-source Java-based package for statistical NLP, document classification, clustering, topic modeling, information extraction, and other machine learning methods that can be applied to text [36]. The Mallet topic model package includes an extremely fast and highly scalable implementation of Gibbs sampling, efficient methods for document-topic hyper parameter optimization, and tools for inferring the topics of new documents given trained models. In this paper, we train the biHTM topic tree on complete Medline abstracts based on the modified Mallet toolkit.

C. PREPROCESSING

The abstract texts of the complete Medline abstracts are extracted from the normalized xml files with the simple API for xml (SAX) parser. Then, we delete 127 canonical stopwords and perform stemming for the abstract texts using the Stanford Natural Language Toolkit [37]. Next, the abstract texts of the complete Medline abstracts are divided into a set of doculets with a 3-sentence sliding window. The Medline abstracts produce a set of over 109 million three-sentence doculets. Finally, word vectors and vocabularies are generated for biHTM training with Mallet.

D. PARAMETERS

The complete Medline abstract doculets contain more than 20 GB of information, and training with a modified Mallet toolkit on a common computer is difficult due to memory limitations. We train the biHTM topic tree for the complete Medline abstracts with a FAT node server with 96 Intel processors, 3 TB of internal memory and a high-capacity shared parallel-access disk.

E. RESULTS

Table 1 shows the key adaptive parameters of the first 3-level biHTM tree nodes for the complete Medline abstracts. These parameters are introduced in section III, where μ_1, μ_2, σ_1, σ_2, α, and β are the parameters of the Gaussian mixture curve fitting the distribution of doculets in each node. x_1 and x_2 are the appropriate topic proportions used to split the doculets. Most x_2 values are distributed in [0.29, 0.40], and most x_2 values are distributed in [0.58, 0.68].

The generated biHTM topic tree for the complete Medline abstracts is shown in Figure 7. The tree contains 14102 nodes, of which 7052 are leaf nodes. The longest path from root to leaf is 18 levels, while the shortest path is 6 levels, with an average of 14.35 levels. The specific semantics of the topic words will be analyzed in the comparison section. There are some null leaf nodes in the tree, and no doculets are split to these nodes.

The topics of the first 3 levels are listed in the Appendix. Figures 8, 9 and 10 show the learning process in the first 3-level nodes of the biHTM topic tree for the complete Medline abstracts. The distributions of doculets show obvious bimodality in all 3-level topic nodes. A large number of doculets are distributed in [0, 0.2] and [0.8, 1], showing that

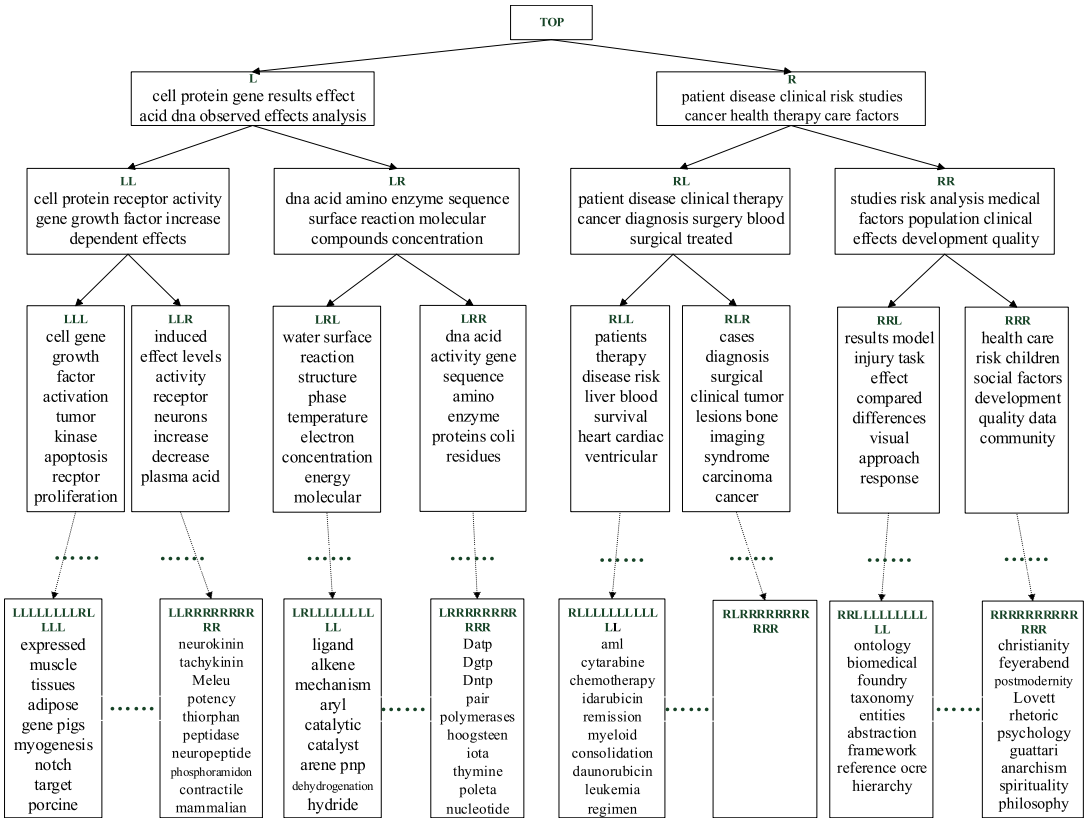


FIGURE 7. The biHTM topic tree for the complete Medline abstracts.

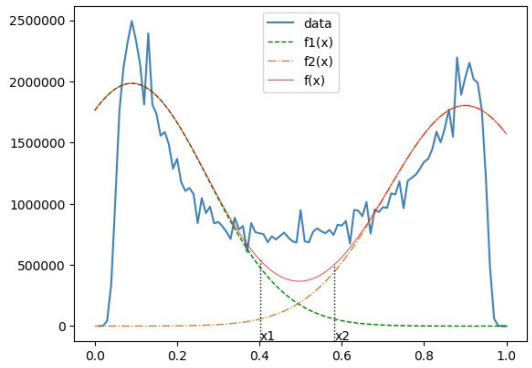


FIGURE 8. The distribution of doculets in the biHTM top level.

most of the doculets are close to 2 topics in each node. In particular, in Figure 10, more doculets in node RL are distributed in [0.4, 0.6]. Two topics in this node are separately related to "therapy and risk" and "disease diagnosis", and the correlation between the two topics is strong.

V. COMPARISON

In this section, we compare biHTM, hLDA, and HLTA in terms of topic interpretability, topic quality and efficiency. As a comparison, Medline abstracts are preprocessed into a set of doculets to build the biHTM topic hierarchy, while hLDA and HLTA use original abstract paragraphs rather than doculets. We train hLDA using the implementation of Mallet, and HLTA using the implementation[3] in [24].

It is very difficult to train the complete Medline abstracts with hLDA. As an alternative, more than 300,000 abstracts are randomly sampled from the complete Medline abstracts for comparison. The depth of the hLDA topic hierarchy is set to 3, the number of topic words is set to 20, and the number of iterations is set to 1000. A specific number of key words in the vocabulary will be extracted first in HLTA, and then the latent topic tree will be built in a bottom-up manner. We train HLTA with the default parameters for the same abstract sample, and the number of key words is set to 10000.

A. INTERPRETABILITY

1) hLDA

Half the words in the topics generated by hLDA in Table 2, such as *patients*, *blood*, *study*, *significantly*, *compared*, *control*, *group*, *levels* and *rate*, are redundant. There are no obvious distinctions between these topics. It is also difficult to give good descriptions of these words. The topics in Table 3 are slightly better than those in Table 2 and include cardiology, nervous system, neuroendocrinology, disease with cell proliferation differentiation and apoptosis,

[3] https://github.com/kmpoon/hlta

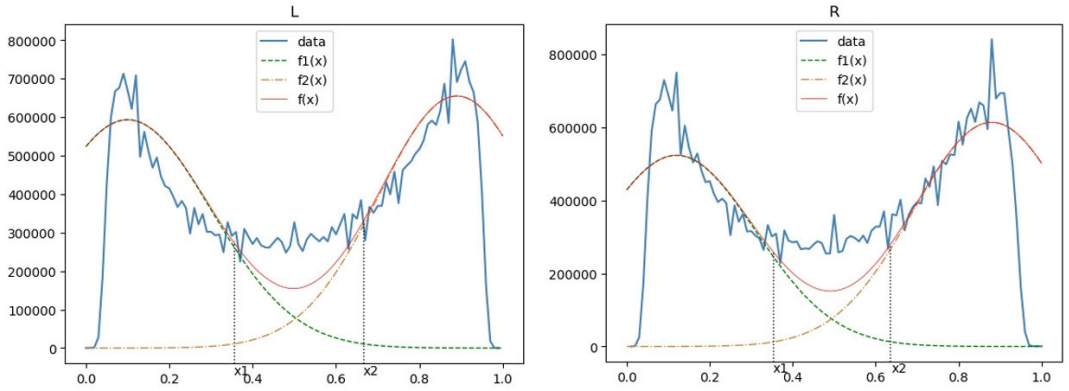


FIGURE 9. The distributions of doculets in the biHTM second level.

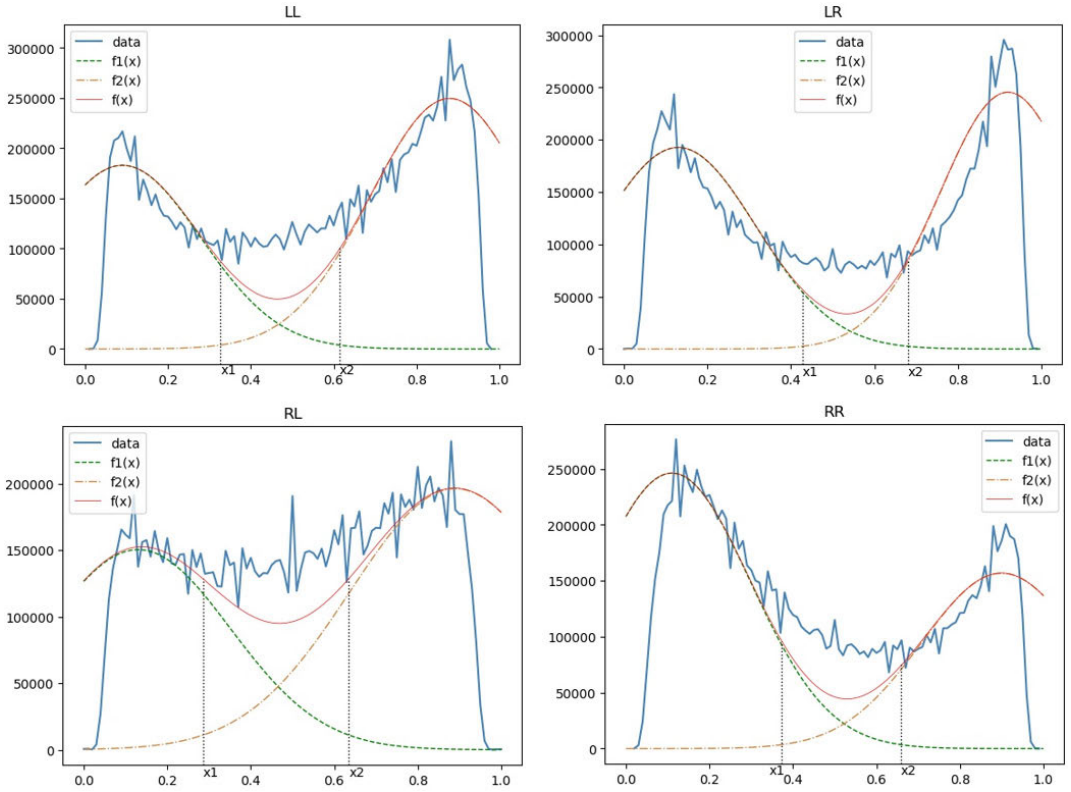


FIGURE 10. The distributions of doculets in the biHTM third level.

microbiology, and DNA. The meanings of topic 2 and topic 3 in Table 3 are very similar, and the relationships between the topics in Table 2 are ambiguous.

2) HLTA

HLTA is different from hLDA because it explores a hierarchy of word co-occurrence relationships. Each topic of HLTA has several key words because a fixed number of observed words could be chosen to train the HLTA topic hierarchy in a bottom-up manner. It can be seen from Table 4 and Table 5 that topics resulting from HLTA are significantly better than topics resulting from hLDA, and there are few repeated words in topic nodes at the same level. Obviously, it can be seen from the Table 4 that topic 1 is about medical R&D and education, topic 2 is about academic work, topic 3 is about respiratory disease, topic 4 is about cell division, topic 5 is about genetics, topic 6 is about cytochemistry, topic 7 is about cancer treatment, topic 8 is about nervous system, topic 9 is about immunology, topic 10 is about alimentary disease, and topic 11 is about medical imaging.

However, topic 1 and topic 2 seem similar. In addition, HLTA extracts a specific number of key words as the

TABLE 2. The partial hLDA topics at the top level.

Node	Topics
topic 1	patients blood increased significantly significant study treatment group compared control levels increase normal days disease pressure rate effects time groups
topic 2	patients treatment group study significant age groups control disease significantly women levels subjects blood compared time higher increased rate effects
topic 7	patients study treatment group significantly significant levels control blood rate increased compared normal response higher time showed age increase observed
topic 10	patients group study compared significant normal increased significantly rate showed groups blood levels results control higher effect found time studied

TABLE 3. The partial hLDA topics in the second level.

Node	Topics
topic1	patients coronary artery ventricular left cases surgery years myocardial patient performed group complications surgical cardiac months treatment risk postoperative aortic
topic2	cells neurons nerve nucleus fibers cell brain cortex muscle rat neuronal expression spinal immunoreactivity area dorsal stimulation axons synaptic rats
topic3	rats receptor mg/kg receptors effect effects binding dopamine release antagonist administration brain activity rat acid secretion response induced agonist doses
topic4	cells cell expression mice human lymphocytes gene protein tumor antigen growth receptor beta antibodies antibody response alpha lines mRNA proliferation
topic6	strains isolates samples water bacteria concentrations species resistance mic degrees tested concentration resistant milk micrograms/ml isolated method growth aureus strain
topic10	binding structure protein residues DNA site enzyme complex peptide amino conformation acid spectra structures form degrees proteins residue NMR sequence

TABLE 4. The HLTA topics at the top level.

Node	Topics
topic1	health care medical practice research education program
topic2	acad natl proc sci coefficient usa experimental-datum
topic3	streptococcus pneumoniae strain haemophilus serotype influenza mycoplasma
topic4	cell monoclonal-antibody inhibit lymphocyte culture inhibition cell-line
topic5	gene sequence dna mrna transcription chromosome encode
topic6	enzyme atp mitochondria mitochondrial atpase dehydrogenase chem
topic7	patient therapy disease treatment treat carcinoma tumor
topic8	neuron nerve nucleus muscle amplitude axon medial
topic9	glutathione gsh glutathione-stransferase glutathione-gsh oxidize oxidant alphatocop
topic10	gastric anf helicobacter-pylori pubertal factor-anf ulcer epilepsy
topic11	conformation circular-dichroism betasheet nmr conformational resonance helix

observation variables, and each word appears in only one branch of the topic hierarchy, HLTA topic hierarchies may be not as comprehensive as hLDA topic hierarchies. In contrast

TABLE 5. The partial HLTA topics at the second level.

Node	Topics
topic 11	health medical care research practice program education
topic 12	antidepressant 8ohdpat 5ht1a-receptor 5ht1a penile tricyclic fluoxetine
topic 13	disabled member disability africa genus phylogenetic family
topic 14	birth-weight gestational-age infant week-gestation neonatal birthweight neonate
topic 21	collect seasonal season winter summer malaria plasmodium
topic 22	acad natl proc sci coefficient usa experimental-datum
topic 23	task memory performance cognitive learning object microscopy
topic 31	streptococcus strain pneumoniae serotype haemophilus bacterium influenza
topic 32	limb hind sms bud hormonereleasing lhrh amputation
topic 33	capd lipoprotein peritoneal-dialysis cholesterol low-density ldl apolipoprotein

TABLE 6. The partial biHTM topics descriptions in levels 1–3.

Node	Topic description	Node	Topic description
L	biological laboratory research	**R**	medical clinical research
LL	cell biology and cell metabolism	**RL**	disease diagnosis and treatment
LLL	disease with cell proliferation differentiation and apoptosis	**RLL**	treatment and risk
LLR	nervous system	**RLR**	disease diagnosis
LR	genetics	**RR**	medical data analysis
LRL	cytochemistry	**RRL**	medical information system
LRR	DNA and protein	**RRR**	influencing factors in treatment

to the topics resulting from hLDA in Table 3, we do not find topics about cardiology and endocrinology in Table 4.

The topics in Table 5 are subtopics of topics 1, 2 and 3 in Table 4, topic 11 to topic 14 are subtopics of topic 1, topic 21 to topic 23 are subtopics of topic 2, and topic 31 to topic 33 are subtopics of topic 3. There seems to be stronger associations between topic 11 and topic 1, topic 22 and topic 2, topic 31 and topic 3 than others. In the process of generating a topic, HLTA will select one subtopic as the main part, and other subtopics are unimportant background topics. As a result, the parent topic may not contain all subtopic information.

3) biHTM

The biHTM topic descriptions of the top three levels are listed in Table 6. The topic of node L is apparently for biological laboratory research, and the topic of node R is for medical clinical research.

In Table 7, with the key words *cell*, *protein*, *gene*, *DNA*, *receptor* and *observed*, it is apparent that the topic description in node L is for **biological laboratory research**, and with the keywords *patients*, *treatment*, *disease*, *cancer*, *clinical*, *therapy* and *blood*, the topic description in node R is for **medical**

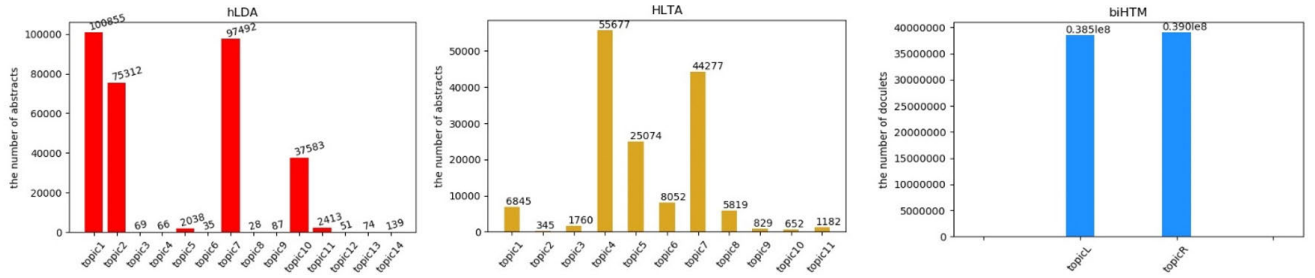


FIGURE 11. The doculet distributions of top node levels in hLDA, biHTM and HLTA.

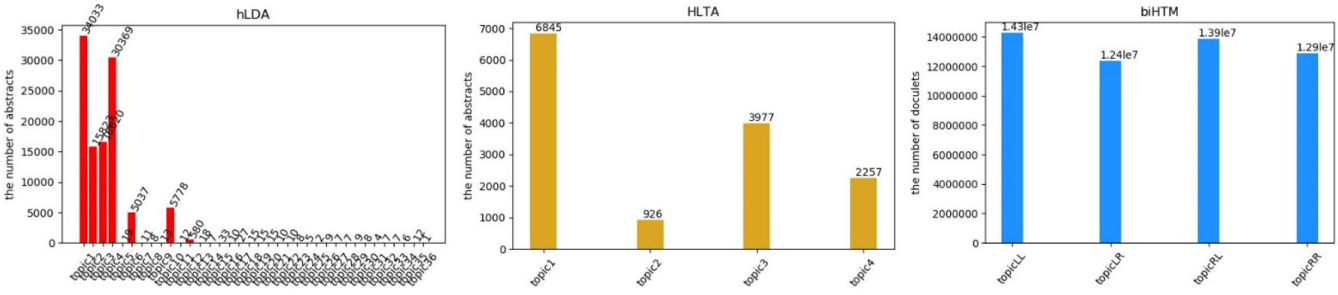


FIGURE 12. The doculet distributions of second level node 1 in hLDA (left), second level node 1 in HLTA (middle) and second level in biHTM (right).

TABLE 7. The two biHTM topics in the top level.

Topic words in node L	Topic words in node R
cells cell protein activity expression induced gene human results levels increased effect acid found high DNA binding showed specific growth observed effects mice receptor proteins type alpha beta analysis role	patients study treatment disease clinical group data results risk cases studies based patient cancer time health years age high significant therapy blood factors related control compared case children analysis care

TABLE 8. Coherence of hLDA, HLTA and biHTM.

Model	NPMI		PMI	
	Average	Median	Average	Median
nLDA	0.044	0.024	0.172	0.180
HLTA	0.114	0.105	0.321	0.198
biHTM	0.052	0.056	0.257	0.157

clinical research, coinciding with most practitioners' divisions of biomedical-related articles. The obvious bimodality reflects the fact that there are two types of articles in the biomedical science literature. The above result indicates that the topics generated by the biHTM have good interpretability. In addition, the relationships among hierarchical topics in the biHTM are clear and reasonable, as the child nodes of medical clinical research in node R are disease diagnosis and treatment and medical data analysis.

It can be seen from Table 6 that the topics resulting from biHTM basically cover the contents in Table 3 and Table 4. Compared with HLTA, the topics generated by biHTM are more comprehensive, and topics resulting from biHTM are significantly more interpretable than topics resulting from hLDA. Although some topics resulting from all three models are similar in meaning, the resultant topic hierarchy obtained by biHTM is much more reasonable.

B. TOPIC QUALITY

Until now, there has been no good metric for measuring the quality of topic hierarchies. We use topic balance and coherence to measure quality in this paper.

The numbers of doculets in different hLDA topic nodes vary substantially, as shown in Figure 11 and Figure 12, and too many hLDA topic nodes are meaningless, with few doculets in most levels. The distribution of documents in the HLTA topic nodes is more even than that in hLDA. In contrast, the distributions of doculets in the biHTM topic nodes are uniform, and the number of doculets in different biHTM topic nodes varies, but not substantially.

Some studies have attempted to automatically and quantitatively estimate topic quality. Newman *et al.* introduced the concept of topic coherence and proposed an automatic method for estimating topic coherence based on pairwise pointwise mutual information (PMI) between the topic words [38]. Han *et al.* experimented with normalized pointwise mutual information (NPMI) [2].

$$\mathrm{PMI}\left(w_i\right) = \sum_{j}^{N-1} \log \frac{P\left(w_i, w_j\right)}{P\left(w_i\right) P\left(w_j\right)} \tag{22}$$

$$\mathrm{NPMI}\left(w_i\right) = \sum_{j}^{N-1} \frac{\log \frac{P(w_i, w_j)}{P(w_i) P(w_j)}}{-\log P\left(w_i, w_j\right)} \tag{23}$$

The PMI and NPMI scores of hLDA, HLTA and biHTM in the second level are calculated using equations (22) and (23),

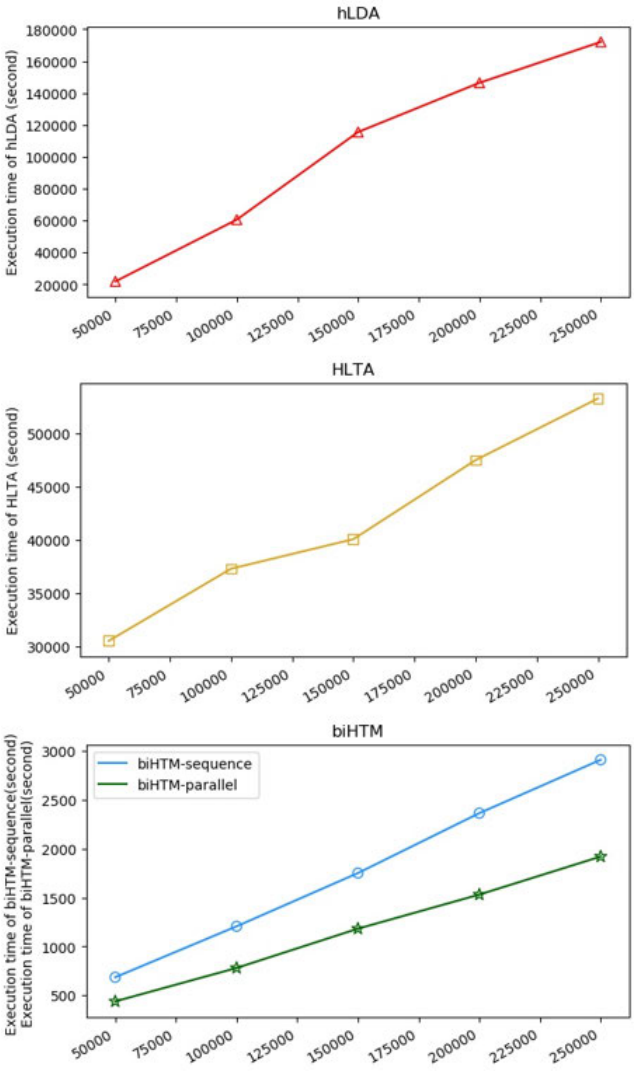


FIGURE 13. Execution times of hLDA, HLTA and biHTM with different numbers of Medline abstracts.

respectively, and the results are shown in Table 8. Because of the large word space and sparsity, none of the scores seem large. HLTA achieves the highest scores due to its model patterns of word co-occurrence using a hierarchical latent tree model. The biHTM achieves higher scores than hLDA, with the exception of the median PMI score. It is obvious that the biHTM topic hierarchy has relatively good coherence.

C. EFFICIENCY

Efficiency is analyzed by employing Medline abstracts of different sizes. In this experiment, the flat topic estimations in biHTM are executed in two different ways: in sequence and in parallel. In sequence, each LDA estimation step is executed separately, and there is only one executing LDA process. In parallel, LDA estimations can be executed concurrently in the fat node server with the maximum amount of concurrency. Figure 13 shows that all execution times increase when the

TABLE 9. The biHTM topics for the complete medline abstracts of level 1–3.

Node	Topic
L	cell protein activity expression induced gene human results levels increased effect acid found high DNA binding showed specific growth observed effects mice receptor proteins type alpha beta analysis role
LL	pubmed cell expression protein induced activity human increased levels receptor effect mice alpha gene results effects response activation dependent role beta growth factor specific significantly study binding increase treatment
LLL	cells pubmed cell expression protein human gene factor growth activation mice role tumor induced specific proteins kinase signaling apoptosis results activity cancer receptor anti beta proliferation genes expressed study type
LLR	pubmed induced increased levels effect activity rats receptor increase effects significantly alpha neurons rat mice release treatment decreased receptors reduced response dependent results plasma beta observed cells acid control treated
LR	pubmed acid DNA protein high binding analysis results found structure species activity amino enzyme sequence surface molecular proteins gene showed reaction based method water complex study process observed compounds concentration
LRL	pubmed surface water high reaction structure process method results phase temperature electron properties formation concentration observed state based solution energy time molecular complex rate obtained complexes low model acid metal
LRR	pubmed protein DNA acid activity species gene sequence amino enzyme binding proteins analysis found showed strains genes beta site strain identified isolated type acids specific residues sequences study high coli
R	patients study treatment disease clinical group data results risk cases studies based patient cancer time health years age high significant therapy blood factors related control compared case children analysis care
RL	pubmed patients treatment disease clinical cases study patient group years therapy cancer diagnosis case year risk surgery blood months age performed rate significant surgical pain acute high treated compared results
RLL	pubmed cases patients case patient diagnosis treatment clinical cancer year surgical liver surgery report tumor lesions bone present performed disease rare

TABLE 9. ***(Continued.)*** **The biHTM topics for the complete medline abstracts of level 1–3.**

	reported syndrome infection presented imaging primary carcinoma results study
RRL	patients pubmed disease group treatment study therapy risk years blood significant rate age months artery compared pressure injury significantly survival acute increased coronary time clinical heart left groups high cardiac
RR	pubmed study health data care based results studies research children risk analysis time model medical related information factors system population clinical quality high women effects development age control important social
RRL	pubmed data results model based time analysis study performance method test effects models studies information system high methods control task subjects effect found compared differences present cognitive visual showed response
RRR	pubmed health care study research patients risk clinical children medical social treatment women factors patient quality development years based life practice related studies data article support students population community important

size of the corpus increases; however, the execution times of hLDA and HLTA are one or two orders of magnitude larger than that of biHTM, as shown in Figure 13. It seems that the execution time of HLTA is smaller than that of hLDA. In addition, if flat topic estimations in biHTM are executed in parallel at all levels, then the execution time of biHTM could decrease by 1/3. Therefore, biHTM is much more efficient than hLDA and HLTA.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a top-down binary hierarchical topic model, called biHTM, for mining biomedical literature. We preprocess the corpus to obtain doculets and then iteratively process LDA estimations for these doculets and their 2 sub-doculet sets to build a biHTM topic tree with adaptive training parameters. The biHTM is a heuristic generative method rather than a probabilistic generative method, unlike most other existing hierarchical topic modeling methods. The biHTM can learn the topic hierarchy very quickly without involving any latent variables. The results show that the topic hierarchy generated by biHTM has good interpretability and relatively high quality. In addition, the distinction between the two sibling topic nodes of the biHTM topic tree is clear, and the relationship between the parent node and child nodes is reasonable. Compared with hLDA and HLTA, biHTM is obviously more suitable for addressing with the massive amount of biomedical literature.

A topic tree with good interpretability will provide strong support for future work, such as biomedical document classification, information extraction, and information retrieval.

The biHTM is proposed to process biomedical literature, such as Medline, in this paper, and it is not always applicable for other large-scale text sets. Numerous experiments can be carried out to confirm this hypothesis in the future. In addition, biHTM can be further extended; for instance, the number of topics generated in each node could be adaptively determined based on the doculets. In addition, we are working with medical experts to analyze the topic tree for the Medline abstracts in more detail, and we will implement a novel text search engine for biomedical documents based on the biHTM topic tree. We will propose these studies in subsequent articles.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their valuable comments and suggestions.

APPENDIX: THE BIHTM TOPICS FOR THE COMPLETE MEDLINE ABSTRACTS OF LEVEL 1-3

See Table 9.

REFERENCES

[1] P. Yan, W. Jin, and K. Jha, "Discovering semantic relationships between concepts from MEDLINE," in *Proc. IEEE 10th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2016, pp. 370–373.

[2] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Gothenburg, Sweden, 2014, pp. 530–539.

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

[4] J. Chen, J. Zhu, J. Lu, and S. Liu, "Scalable training of hierarchical topic models," in *Proc. 44th Int. Conf. Very Large Data Bases (VLDB)*, Mar. 2018, pp. 826–839.

[5] 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, pp. 1–30, Jan. 2010.

[6] D. M. Blei, M. I. Jordan, T. L. Griffiths, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," presented at the 16th Int. Conf. Neural Inf. Process. Syst., Whistler, BC, Canada, 2003.

[7] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested hierarchical Dirichlet processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 256–270, Feb. 2015.

[8] T. Liu, N. L. Zhang, and P. Chen, "Hierarchical latent tree analysis for topic detection," in *Proc. Mach. Learn. Knowl. Discovery Databases-Eur. Conf. (ECML PKDD)*, 2014, pp. 256–272.

[9] 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. (EMNLP-CoNLL)*, 2012, pp. 800–809.

[10] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, "A hierarchical topic modelling approach for tweet clustering," in *Proc. 9th Int. Conf. Social Informat. (SocInfo)*, Oxford, U.K., 2017, pp. 378–390.

[11] 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.

[12] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, vol. 51, 1999, pp. 50–57.

[13] W. Buntine and A. Jakulin, "Applying discrete PCA in data analysis," in *Proc. 20th Conf. Uncertainty Artif. Intell.*, 2004, pp. 59–66.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5228–5235, Apr. 2004.

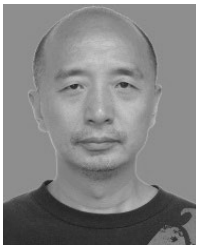
[15] D. Newman, S. Karimi, and L. Cavedon, "Using topic models to interpret MEDLINE's medical subject headings," in *Proc. 22nd Australas. Joint Conf. Artif. Intell.*, 2009, pp. 270–279.

[16] B. Zheng, D. C. McLean, and X. Lu, "Identifying biological concepts from a protein-related corpus with a probabilistic topic model," *BMC Bioinf.*, vol. 7, no. 1, p. 58, Dec. 2006.

[17] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with pachinko allocation," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 633–640.
[18] A. M. Dai and A. J. Storkey, "The supervised hierarchical Dirichlet process," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 243–255, Feb. 2015.
[19] W. Wang, H. Xu, Y. Weiwei, and H. Xiaoqiu, "Constrained-hLDA for topic discovery in Chinese microblogs," in *Proc. 18th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, Tainan, Taiwan, 2014, pp. 608–619.
[20] Y. Xu, J. Yin, J. Huang, and Y. Yin, "Hierarchical topic modeling with automatic knowledge mining," *Expert Syst. Appl.*, vol. 103, pp. 106–117, Aug. 2018.
[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.
[22] X. Zou, Y. Zhu, J. Feng, J. Lu, and X. Li, "A novel hierarchical topic model for horizontal topic expansion with observed label information," *IEEE Access*, vol. 7, pp. 184242–184253, 2019.
[23] 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.
[24] P. Chen, N. L. Zhang, T. Liu, L. K. M. Poon, Z. Chen, and F. Khawar, "Latent tree models for hierarchical topic detection," *Artif. Intell.*, vol. 250, pp. 105–124, Sep. 2017.
[25] D. Stück, H. T. Hallgrímsson, G. Ver Steeg, A. Epasto, and L. Foschini, "The spread of physical activity through social networks," in *Proc. 26th Int. Conf. World Wide Web (WWW)*, Perth, Australia, 2017, pp. 519–528.
[26] G. V. Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 577–585.
[27] A. Ahmed, L. Hong, and A. J. Smola, "Nested Chinese restaurant franchise process: Applications to user tracking and document modeling," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, USA, 2013, pp. 1426–1434.
[28] M. Chen, "Research on key technology of Web hierarchical topic detection and evolution based on behaviour tracking analysis," *Int. J. Comput. Commun. Control*, vol. 14, no. 3, pp. 311–328, 2019.
[29] L. Wang, L. La, and Z. Wang, "A three stage method for inter-topic correlation analysis in social networks," *J. Nonlinear Convex Anal.*, vol. 20, no. 7, pp. 1353–1364, 2019.
[30] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PLoS ONE*, vol. 6, no. 3, 2011, Art. no. e18029.
[31] Y. Wang, L. Huang, S. Guo, L. Gong, and T. Bai, "A novel MEDLINE topic indexing method using image presentation," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 130–137, Jan. 2019.
[32] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018.
[33] R.-C. Chen, R. Swanson, and A. S. Gordon, "An adaptation of topic modeling to sentences," 2016, *arXiv:1607.05818*. [Online]. Available: http://arxiv.org/abs/1607.05818
[34] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. Mccallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Edinburgh, Scotland, 2011, pp. 262–272.
[35] T. C. O'Haver, *A Pragmatic Introduction to Signal Processing: With Applications in Scientific Measurement*. College Park, MD, USA: CreateSpace Independent Publishing Platform, 2017.
[36] A. K. McCallum. (2002). *MALLET: A Machine Learning for Language Toolkit*. [Online]. Available: http://mallet.cs.umass.edu
[37] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Baltimore, MD, USA, 2014, pp. 55–60.
[38] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL HLT)*, Los Angeles, CA, USA, 2010, pp. 100–108.

XIAOGUANG LIN received the M.S. degree from the University of Chinese Academy of Sciences, China, in 2009, where he is currently pursuing the Ph.D. degree in computer science and technology. His current research interests include text mining, topic model, and biomedical text processing.

MINGXUAN LIU is currently pursuing the M.S. degree with the University of Chinese Academy of Sciences. His research interests include natural language processing and biomedical data analysis.



JU ZHANG received the Ph.D. degree from the University of Rochester, USA, in 1994. He is currently a Professor with the University of Chinese Academy of Sciences. His research interests include natural language processing, deep learning, text mining, and biomedical data processing.