EHLLDA: A Supervised Hierarchical Topic Model

Xian-Ling Mao[1], Yixuan Xiao[1], Qiang Zhou[1], Jun Wang[2], and Heyan Huang[1]

[1]Department of Computer Science and Technology, Beijing Institute of Technology
[2]Institute of Biz Big Data, Sogou Inc.
[1]{maoxl,1120121905,qzhou,hhy63}@bit.edu.cn
[2]wangjunbj7526@sogou-inc.com

Abstract. In this paper, we consider the problem of modeling hierarchical labeled data – such as Web pages and their placement in hierarchical directories. The state-of-the-art model, hierarchical Labeled LDA (hLLDA), assumes that each child of a non-leaf label has equal importance, and that a document in the corpus cannot locate in a non-leaf node. However, in most cases, these assumptions do not meet the actual situation. Thus, in this paper, we introduce a supervised hierarchical topic models: *Extended Hierarchical Labeled Latent Dirichlet Allocation* (EHLLDA), which aim to relax the assumptions of hLLDA by incorporating prior information of labels into hLLDA. The experimental results show that the perplexity performance of EHLLDA is always better than that of LLDA and hLLDA on all four datasets; and our proposed model is also superior to hLLDA in terms of p@n.

Keywords: Topic modeling; Supervised Learning; Hierarchical Topic Modeling

1 Introduction

A number of topic models have been developed for the data without labels [11, 26, 28], and the data with non-hierarchical labels [24, 3, 23]. For the data with hierarchical labels, like webpages and their corresponding hierarchical directories, to the best of our knowledge, the *hierarchical Labeled Latent Dirichlet Allocation* (hLLDA) [19] is the only topic model proposed to model this kind of data. The generative process of hLLDA is: (1) choose a random path c_d for a document d among all the paths in the hierarchical labeled tree; (2) draw a proportion over the labels in path c_d; (3) each of the N words in d is selected from one of the topics (labels). Note that hLLDA takes each label as a topic, i.e. a distribution over vocabulary, thus hLLDA needs to learn a distribution for each label. In this paper, a "label" means a character string or a distribution over vocabulary, which can be distinguished in different context. From the generative process, hLLDA has two latent assumptions: (i) hLLDA treats each child of a non-leaf label equally (See step (1)); (ii) hLLDA also assumes that each document in the corpus must have a leaf label, i.e. each document cannot locate in the non-leaf node in the hierarchy of labels. However, in most cases, these assumptions do not meet the actual situation. For assumption (i), often each child has different importance. For example, in Yahoo! Answer, the number of questions is different for different sub-categories of a category, which shows the importance of labels is different. For assumption (ii), documents often

locate in intermediate layers. For example, in Yahoo! Answer, the categories in intermediate layers often have questions, which shows that documents can locate in non-leaf nodes. In this paper, we extend hLLDA to a model named *Extended Hierarchical Labeled LDA* (EHLLDA) by taking advantage of prior information of labels and relaxing assumptions.

We demonstrate the effectiveness of the proposed model on large, real-world datasets in the question answering and website category domains. We also observe that prior information is very valuable when incorporated into topic learning.

2 Extended Hierarchical Labeled LDA

In this paper, we introduce a supervised hierarchical topic model, i.e., the *Extended Hierarchical Labeled LDA* (EHLLDA). EHLLDA is a probabilistic graphical model that describes a process for generating a hierarchical labeled document collection. Like the hLLDA, EHLLDA models each document as a mixture of underlying topics and generates each word from one topic; meanwhile EHLLDA incorporates supervision by simply constraining the topic model to use only those topics that correspond to a document's (observed) hierarchical labels. Unlike hLLDA, EHLLDA incorporates prior information of labels by capturing the relation between a parent label and its child labels, i.e., the relation between a super-topic and its sub-topics.

The graphical model of EHLLDA is depicted in Figure 1. Each label in EHLLDA has its corresponding topic. The model can be viewed in terms of a generative process that first generates c_d labels from the hierarchy of labels for a documnet d, and then draw a proportion over the c_d labels, and finally each of the N words in d is selected from one of the c_d topics (labels).

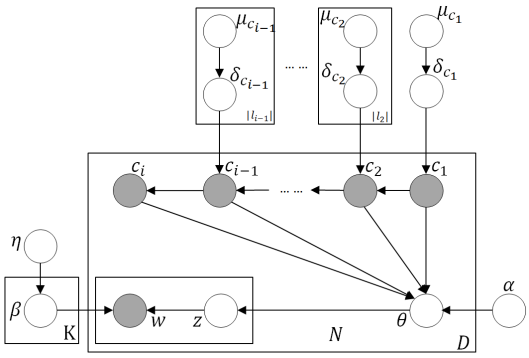

Fig. 1. The graphical model representation of the Extended Hierarchical Labeled LDA.

In the model, N is the number of words in a document, D is the total number of documents in a collection, K is the number of labels in the hierarchy, L is the height of hierarchy of labels, c_i is an observed node in the i^{th} level in the hierarchical labeled

tree for a document, $\boldsymbol{c_d} = \{c_1, c_2, ..., c_{|c_d|}\}$ be the labels for a document d, l_i be the set of labels in the i^{th} level in the hierarchy of labels. η, α and μ_{c_i} are dirichlet prior parameters, β_k is a distribution over words, θ is a document-specific distribution over topics, δ_{c_i} is a multinomial distribution over observed sub-topics of topic c_i, w is an observed word, z is the topic assigned to w, $Dir_k(.)$ is a k-dimensional Dirichlet distribution, $Mult(.|.)$ is a multinomial distribution, γ is a Multi-nomial distribution over paths in the tree, and V is the size of vocabulary.

A EHLLDA model assumes the following generative process for a document and its hierarchical labels ($\boldsymbol{w_d}; \boldsymbol{c_d}$):

1. For each topic $k \in \{1, ..., K\}$.
   (a) Generate $\beta_k = (\beta_{k,1}, ..., \beta_{k,V})^T \sim Dir(.|\eta)$
2. For each level l_j, $j \in \{2, ..., L-1\}$:
   (a) For each node c_i in l_j, draw $\delta_{c_i} \sim Dir(.|\mu_{c_i})$
3. For each document d:
   (a) For each level l_i, $i \in \{2, ..., L\}$:
      i. Draw $c_i \sim Mult(.|\delta_{c_i})$
      ii. If c_i is an "exit" node, goto (b)
   (b) Draw a distribution over the nodes in the set $\boldsymbol{c_d}$, $\theta_d \sim Dir(.|\alpha, \boldsymbol{c_d})$
   (c) For each $i \in \{1, ..., N_d\}$:
      i. Generate $z_i \sim Mult(.|\theta_d)$
      ii. Generate $w_i \sim Mult(.|\beta_{z_i})$

Specifically, we associate with each label c in the hierarchy of labels a document-specific dirichlet distribution with dimensionality equal to $N_c + 1$, where N_c is the number of children of the label node c. This distribution allows us to traverse the hierarchy of labels and exit at any node in the hierarchy of labels — given that we are at a label node c_i, there are N_{c_i} child labels to choose from and an additional option to choose an "exit" child to exit the labeled tree at label node c_i. We start our walk through the hierarchy of labels at the root node and select a node from its children. We repeat this process until we reach an exit node. A word is generated from one of the topics from the root to the parent of the exit node. We illustrate an example of the hierarchical labeled tree in Figure 2 for six documents. It shows the paths of six documents through the hierarchy of labels. The solid lines connect each node to the sub-nodes. The shaded circles stand for observed labels, and black circles stand for "exit" node. For example, the 5^{th} document, it first chooses label *A1* according to a probability distribution, then chooses *A3*, finally chooses *A5*. *A5* is a "exit" node, thus the 5^{th} document has labels: *A1* and *A3*; meanwhile the 5^{th} document is generated by topics: *A1* and *A3*. Here, we can see that the 5^{th} document has located in a non-leaf label, which has relaxed the assumption of hLLDA. In addition, from the generative process, each label has assigned a different choosing probability, which relaxes another assumption of hLLDA.

3 Parameter Estimation

3.1 Learning and Inference

In Figure 1, the labels for a document are observed, so θ and δ_{c_i} are d-separated from the rest of the model given labels $\boldsymbol{c_d}$. Therefore the learning and inference for EHLLDA are similar to traditional LDA, can be solved using collapsed Gibbs sampling.

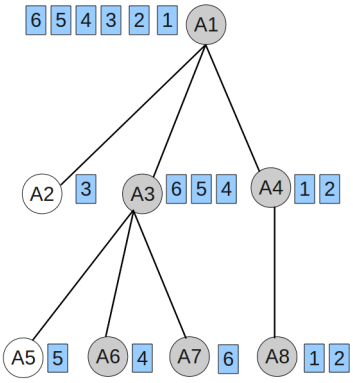

Fig. 2. An example of the hierarchical labeled tree for six documents. It shows the paths of six documents through the hierarchy of labels. The solid lines connect each node to the sub-nodes. The shaded circles stand for observed labels, and black circles stand for "exit" node.

For each document, the topics used for inference are those found in the set of labels from the root to the "exit" node in the hierarchy of labels. Once the target labels $\boldsymbol{c_d}$ is known, the model is reduced to LDA over the set of topics comprising $\boldsymbol{c_d}$. Although the joint distribution $p(\theta, \mathbf{z}, \mathbf{w}|\boldsymbol{c_d})$ is intractable [5], individual word-level assignments can be obtained by collapsed Gibbs-sampling [11]. In collapsed Gibbs-sampling, a Markov chain is constructed to converge to the target distribution, and samples are then taken from that Markov chain. Each state of the chain is an assignment of values to the variables being sampled. To apply this algorithm we need the full conditional distribution $p(z_i = j|\boldsymbol{z_{-i}}, \mathbf{w}, \boldsymbol{c_d})$. Specifically, the probability of assigning w_i, the i^{th} word in document d, to the j^{th} topic in the set $\boldsymbol{c_d}$, conditioned on all other word assignments $\boldsymbol{z_{-i}}$, is given by:

$$p(z_i = j|\boldsymbol{z_{-i}}, \boldsymbol{w}, \boldsymbol{c_d}) \propto \frac{n^{w_i}_{-i,j} + \eta}{V(\eta + 1)} \times \frac{n^{d}_{-i,j} + \alpha}{|\boldsymbol{c_d}|(\alpha + 1)} \tag{1}$$

where $n^{d}_{-i,j}$ is the frequency of words from document d assigned to topic j other than word i, $n^{w_i}_{-i,j}$ is the frequency of word w_i in topic j, that does not include the current assignment z_i, η and α are Dirichlet prior parameters for the topics and topic word multinomials respectively, and V is the size of vocabulary.

Having obtained the full conditional distribution, the Gibbs sampling algorithm is then straightforward. The z_i variables are initialized to determine the initial state of the Markov chain. The chain then runs for a number of iterations, each time finding a new state by sampling each z_i from the distribution specified by Equation (1). After obtaining individual word assignments $\mathbf{z}$, we can estimate the topic multinomials and the per-document mixing proportions. Specifically, the topic multinomials are estimated as

$$\beta_{\boldsymbol{c_d}[j],i} = p(w_i|z_{\boldsymbol{c_d}[j]}) = \frac{\eta + n^{w_i}_{z_{\boldsymbol{c_d}}[j]}}{V\eta + \sum n^{\cdot}_{z_{\boldsymbol{c_d}}[j]}} \tag{2}$$

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

$$\theta_{d,j} = \frac{\alpha + n^d_{.,j}}{|\boldsymbol{c_d}|\alpha + n^d}, j \in 1, ..., |\boldsymbol{c_d}| \tag{3}$$

where $\boldsymbol{c_d}[j]$ means the j^{th} topic in $\boldsymbol{c_d}$. Although the equations above look exactly the same as those of LDA, there is an important distinction in that, the target topic j is restricted to belong to the set of labels in $\boldsymbol{c_d}$.

Dirichlet-multinomial parameter estimation For EHLLDA model, except for $\boldsymbol{z}, \boldsymbol{\theta}$, we have to estimate μ_{c_i}, which are the parameters of Dirichlet-multinomial distribution (Polya distribution). It is a compound distribution where δ_{c_i} is drawn from a Dirichlet $Dir(.|\mu_{c_i})$ and a sample of discrete outcomes **x** is drawn from the multinomial with the probability vector δ_{c_i}. Let n_k be the number of times, and the outcome is k. Then the resulting distribution over **x**, a vector of outcomes, is given as:

$$p(\mathbf{x}|\mu_{c_i}) = \int_{\delta_{c_i}} p(\mathbf{x}|\delta_{c_i}) p(\delta_{c_i}|\mu_{c_i}) d\delta_{c_i} \tag{4}$$

$$= \frac{\Gamma(\sum_k \mu_{c_i k})}{\Gamma(\sum_k n_k + \mu_{c_i k})} \prod_k \frac{\Gamma(n_k + \mu_{c_i k})}{\Gamma(\mu_{c_i k})} \tag{5}$$

This distribution is also parameterized by $\mu_{c_i k}$, which can be estimated from a training set of count vectors: D = $\{x_1, ..., x_N\}$. The likelihood is

$$p(D|\mu_{c_i}) = \prod_j p(x_i|\mu_{c_i}) \tag{6}$$

$$= \prod_j \left(\frac{\Gamma(\sum_k \mu_{c_i k})}{\Gamma(\sum_k \mu_{c_i k} + n_j)} \prod_k \frac{\Gamma(n_{jk} + \mu_{c_i k})}{\Gamma(\mu_{c_i k})} \right) \tag{7}$$

We apply fixed-point iteration [17] to maximize the gradience of $\log p(D|\mu_{c_i})$ as follows:

$$\mu^{new}_{c_i k} = \mu_{c_i k} \frac{\sum_j \Psi(n_{jk} + \mu_{c_i k}) - \Psi(\mu_{c_i k})}{\sum_j \Psi(n_j + \sum_k \mu_{c_i k}) - \Psi(\sum_k \mu_{c_i k})} \tag{8}$$

where Ψ is the digamma function. Through Equation (8), we can obtain the estimation of μ_{c_i}.

4 Experiment

4.1 Dataset

To construct comprehensive datasets for our experiments, we crawled data from two websites. First, we crawled question-answer pairs (QA pairs) of two top categories of Yahoo! Answers: *Computers* & *Internet* and *Health*. We refer to the data from the category *Computers* & *Internet* as *Y_Comp*, and the data from the category *Health* as *Y_Hlth*. In addition, we first crawled two categories of Open Directory Project (ODP)[1]: *Home* and *Health*. Then, we removed all categories whose numbers of Web sites are less than

[1] http://dmoz.org/

3. Finally, for each Web site in the categories, we submited its url to Google and used the words in the snippet and title of the first returned result to extend the summary of the Web site. We denote the data from the category *Home* dataset as *O_Home*, and the data from the category *Health* as *O_Hlth*. The statistics of all datasets are summarized in Table 1.

Because hLLDA cannot process the situation in which there are documents in non-leaf nodes, we treat the corresponding label of each document as its path in hLLDA to ensure fairness.

Table 1. The statistics of the datasets.

Datasets	#labels	#paths	Max level	#docs
Y_Comp	27	23	4	3,203,793
Y_Hlth	28	24	4	4,122,983
O_Hlth	6695	6505	10	54939
O_Home	2432	2364	9	24254

4.2 Performance of Proposed Topic Models

In the area of topic modeling, there are usually three methods to evaluate the proposed model: (i) Case study; (ii) Perplexity; (iii) Evaluated indirectly in the third-party application. Here, we will first observe a training result from the proposed model, and then evaluate how well the proposed model describes a dataset in terms of *perplexity*, and finally evaluate the ranking quality of the model, comparing with hLLDA.

Case Study With topic modeling, the top associated words of topics can be used as good descriptors for labels in a hierarchy. In Figure 3, we show an example of a path of categories "/Computers & Internet/Hardware/Laptops & Notebooks" from the *Y_Comp* dataset. The topics are the results of EHLLDA with 1500 Gibbs sampling iterations, and symmetric priors $\alpha = 0.01$, $\eta = 0.01$.

We made two major observations from the example: (i) topic words for higher level categories are more general than lower level ones. For example, words like "open", "files", "click" and "installed" are associated with the top *"Computer & Internet"*, while more specific words like "screen", "usb" and "keyboard" are associated with lower category *"Hardware"*. This shows that EHLLDA is capable of capturing the hierarchical topic structure of the dataset. (ii) The parent-child relation is reflected by the phenomenon of inherence: some topic words appear in categories that are in a path. For example, "screen" and "disk" appear in both *"Computers & Internet"* and *"Hardware"* which are considered a pair of parent-child categories. Note that the two common words have different importance in the two categories, which further verifies that parent-child categories are related but different. These observations further confirm that EHLLDA is a hierarchy structure aware topic model.

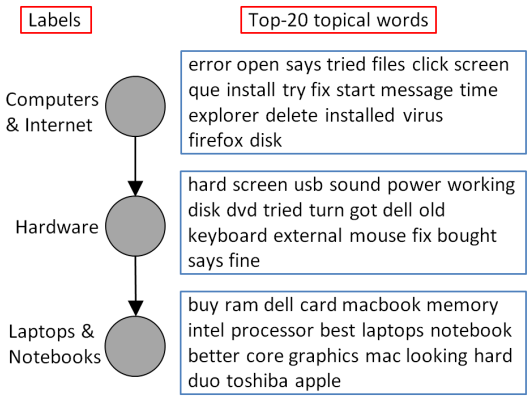

Fig. 3. A topical hierarchy learned with EHLLDA for the path "/Computers & Internet/Hardware/Laptops & Notebooks" in Yahoo! Answer dataset; the top 20 words are shown for each topic. Labels are shown on the left side, and the topical words of each label are shown on the right side.

Measure by Perplexity A good supervised hierarchical topic model should be able to generalize to unseen data. To measure the prediction ability of our models, we computed the perplexity of the given categories under $p(c_d|d)$ for each document d in the test sets [5]. The perplexity of M test documents is calculated as:

$$perplexity(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\sum_{m=1}^{N_d}\log p(w_{dm})}{\sum_{d=1}^{M}N_d}\right\} \tag{9}$$

where D_{test} is the test collection of M documents, N_d is document length of document d and w_{dm} is the m^{th} word in document d. We trained LLDA [21], hLLDA [19] and EHLLDA, on all four datasets to compare the prediction performance of these models. LLDA is a state-of-the-art supervised non-hierarchical topic model, which does not consider the relation between labels. hLLDA is a state-of-the-art supervised hierarchical topic model, intending to model the relation between labels. Our model, EHLLDA, will compare with LLDA and hLLDA. We keep 80% of the data collection as the training set and use the remaining collection as the held-out test set. We build the models based on the training set with 1500 Gibbs sampling iterations, and symmetric priors α = 0.01, η = 0.01, and compute the preplexity of the test set to evaluate the models. Thus, our goal is to achieve high likelihood on a held-out test set.

Table 2 shows the perplexity of each model. From the table, we can see that the perplexities of all supervised hierarchical topic models, i.e., hLLDA and EHLLDA, are lower than supervised non-hierarchical topic model – LLDA. It shows that the performance of models that consider the relation between labels is better than that without considering the relations between labels. Furthermore, we can also see that the perplexities of EHLLDA are lower than that of hLLDA over all four datasets. The results show that our proposed model can model the supervised hierarchical data better than the state-of-the-art model – hLLDA.

Table 2. Perplexity of the datasets.

Datasets	LLDA	hLLDA	EHLLDA
Y_Comp	33296.4	22952.3	**21848.3**
Y_Hlth	3017.8	2998.7	**2994.5**
O_Hlth	1667721.1	108640.0	**93954.9**
O_Home	1196459.3	116541.8	**95989.3**

Measure by p@n For each test document, we run the comparing systems to predict a ranking of all C possible paths and compare their performance in terms of precision at top n (p@n). hLLDA is used as our baseline algorithm again. For all the datasets and comparing models, we keep 80% of the data in the collection as the training set and the remaining collection as the test set. All models are trained with 1500 Gibbs sampling iterations, and symmetric priors $\alpha = 0.01$, $\eta = 0.01$. The experimental results are shown in Table 3. From the table, we can see that EHLLDA outperforms the baseline method (hLLDA) significantly on all four datasets. The improvement is significant by t-test with 95% significance. This suggests that EHLLDA is better at modeling the topics of the documents thus leads to better ranking results. From the table, the p@n values of hLLDA and EHLLDA over O_Home and O_Hlth are very low. This is because there are too many paths in these two datasets, thus it's hard to discriminate these paths for baseline and proposed algorithms. However, since our aim is to verify the proposed model is better than the baseline by the ranking problem, not to research the ranking problem itself, low p@n values don't change our conclusion.

Table 3. Ranking Predictions for each dataset.

Datasets	**Models**	**Measures**			
		P@1	P@2	P@5	P@10
Y_Comp	hLLDA	0.2106	0.2585	0.3688	0.4599
	EHLLDA	**0.2417**	**0.3181**	**0.4552**	**0.5261**
Y_Hlth	hLLDA	0.3655	0.4601	0.5595	0.6103
	EHLLDA	**0.4139**	**0.5027**	**0.5932**	**0.6426**
O_Home	hLLDA	0.0206	0.0289	0.0510	0.0669
	EHLLDA	**0.0283**	**0.0397**	**0.0546**	**0.0809**
O_Hlth	hLLDA	0.0350	0.0462	0.0730	0.0914
	EHLLDA	**0.0418**	**0.0560**	**0.0849**	**0.1113**

5 Related works

Topic model has been widely and successfully applied to blog articles and other text collections to mine topic patterns [4, 5]. There have been many variations of topic models (TM). The existing topic models can be divided into four categories: *Unsupervised*

non-hierarchical topic models, Unsupervised hierarchical topic models, and their corresponding supervised counterparts.

Unsupervised non-hierarchical topic models are widely studied, such as LSA [9], pLSA [12], LDA [5], Hierarchical-concept TM [8, 7], d-BTM [27], Correlated TM [2], TMIO [10] and Concept TM [6, 7] etc. The most famous one is Latent Dirichlet Allocation (LDA). LDA is similar to pLSA, except that in LDA the topic distribution is assumed to have a Dirichlet prior. LDA is a completely unsupervised algorithm that models each document as a mixture of topics. Another famous model that does not only represents topic correlations, but also learns them, is the Correlated Topic Model (CTM). Topics in CTM are not independent; however it is noted that only pairwise correlations are modeled, and the number of parameters in the covariance matrix grows as the square of the number of topics.

However, the models above cannot capture the relation between super and sub topics. To address this problem, many models have been proposed to model the relations, such as Hierarchical LDA (HLDA) [1], Hierarchical Dirichlet processes (HDP) [26], Hierarchical PAM (HPAM) [16] and nHDP [13] etc. The relations are usually in the form of a hierarchy, such as the tree or Directed Acyclic Graph (DAG). HDP is proposed to model the groups of data that have a pre-defined hierarchical structure. HDP can capture topic correlations defined by this type of nested data structure; However, it does not automatically discover such correlations from unstructured data. To handle the large topic space, PAM, which uses a DAG structure, is developed to represent and learn the arbitrary, nested, and possibly sparse topic correlations. In PAM, the concept of topics is extended to be distributions not only over words, but also over other topics.

Although unsupervised topic models are sufficiently expressive to model multiple topics per document, they are inappropriate for labeled corpora because they are unable to incorporate the supervised label set into their learning procedure. Several modifications of LDA to incorporate supervision have been proposed in the literature. Two such models, Supervised LDA [3, 4] and DiscLDA [14] are first proposed to model documents associated only with a single label. Recently, IRTM [20] is proposed to combine the strengths of MNIR and LDA. Another category of models, such as the MM-LDA [22], Author TM [24], TRTM [15], SNT [13], Prior-LDA [25], Dependency-LDA [25], MedLDA [29] and Partially LDA (PLDA) [23] etc., are not constrained to one label per document because they model each document as a bag of words with a bag of labels, with topics for each observation drawn from a shared topic distribution.

None of these models, however, leverage dependency structure, such as parent-child relation, in the label space. HSLDA [18] and hLLDA [19] are proposed to capture the structral relation. HSLDA still needs to decide manually how many topics in a collection, i.e. parameter K. hLLDA takes each label as a topic, i.e. a distribution over vocabulary, thus hLLDA assumes that the number of topics in a labeled collection is the one of labels. From the generative process, hLLDA has two latent assumptions: (i) hLLDA treats each child of a non-leaf label equally; (ii) hLLDA also assumes that each document in the corpus must have a leaf label, i.e. each document cannot locate in the non-leaf node in the hierarchy of labels. However, in most cases, these assumptions do not meet the actual situation. Thus, in this paper, we extend hLLDA to a model

named *Extended Hierarchical Labeled LDA* (EHLLDA) by taking advantage of prior information of labels and relaxing assumptions.

6 Conclusion and Future work

In this paper, we considered the problem of modeling hierarchical labeled data – such as web pages and their placement in hierarchical directories, and product descriptions and catalogs. We proposed a supervised hierarchical topic model, i.e. EHLLDA, which incorporated prior information of paths and relaxed the assumption of hLLDA. The experimental results show that our model is always better than baseline in terms of perplexity and p@n.

In the future, we will continue to explore novel topic models for supervised hierarchical data to further improve the performance; meanwhile we will also apply our supervised hierarchical topic models to other media forms, such as image, to test model's generalization ability and solve related problems in these area.

7 Acknowledgments

The work was supported by National Natural Science Foundation of China (No. 61402036), 863 Program of China (No. 2015AA015404) and 973 Program (No. 2013CB329605).

References

1. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested chinese restaurant process. Advances in neural information processing systems 16, 106 (2004)
2. Blei, D., Lafferty, J.: Correlated topic models. Advances in neural information processing systems 18, 147 (2006)
3. Blei, D., McAuliffe, J.: Supervised topic models. In: Proceeding of the Neural Information Processing Systems(nips) (2007)
4. Blei, D., McAuliffe, J.: Supervised topic models. Arxiv preprint arXiv:1003.0783 (2010)
5. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
6. Chemudugunta, C., Holloway, A., Smyth, P., Steyvers, M.: Modeling documents by combining semantic concepts with unsupervised statistical learning. The Semantic Web-ISWC 2008 pp. 229–244 (2008)
7. Chemudugunta, C., Smyth, P., Steyvers, M.: Combining concept hierarchies and statistical topic models. In: Proceeding of the 17th ACM conference on Information and knowledge management. pp. 1469–1470. ACM (2008)
8. Chemudugunta, C., Smyth, P., Steyvers, M.: Text modeling using unsupervised topic models and concept hierarchies. Arxiv preprint arXiv:0808.0973 (2008)
9. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391–407 (1990)
10. Du, L., Pate, J.K., Johnson, M.: Topic segmentation with an ordering-based topic model. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)

11. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(Suppl 1), 5228 (2004)
12. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, UAI'99. p. 21. Citeseer (1999)
13. Kawamae, N.: Supervised n-gram topic model. In: Proceedings of the 7th ACM international conference on Web search and data mining. pp. 473–482. ACM (2014)
14. Lacoste-Julien, S., Sha, F., Jordan, M.: ndisclda: Discriminative learning for dimensionality reduction and classification. Advances in Neural Information Processing Systems 21 (2008)
15. Ma, Z., Sun, A., Yuan, Q., Cong, G.: A tri-role topic model for domain-specific question answering. In: Proceedings of The Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
16. Mimno, D., Li, W., McCallum, A.: Mixtures of hierarchical topics with pachinko allocation. In: Proceedings of the 24th international conference on Machine learning. pp. 633–640. ACM (2007)
17. Minka, T.: Estimating a dirichlet distribution. Annals of Physics 2000(8), 1–13 (2003)
18. Perotte, A.J., Wood, F., Elhadad, N., Bartlett, N.: Hierarchically supervised latent dirichlet allocation. In: Advances in Neural Information Processing Systems. pp. 2609–2617 (2011)
19. Petinot, Y., McKeown, K., Thadani, K.: A hierarchical model of web summaries. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 670–675. Association for Computational Linguistics (2011)
20. Rabinovich, M., Blei, D.: The inverse regression topic model. In: Proceedings of The 31st International Conference on Machine Learning. pp. 199–207 (2014)
21. Ramage, D., Hall, D., Nallapati, R., Manning, C.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. pp. 248–256. Association for Computational Linguistics (2009)
22. Ramage, D., Heymann, P., Manning, C., Garcia-Molina, H.: Clustering the tagged web. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. pp. 54–63. ACM (2009)
23. Ramage, D., Manning, C., Dumais, S.: Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 457–465. ACM (2011)
24. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 487–494. AUAI Press (2004)
25. Rubin, T., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. Arxiv preprint arXiv:1107.2462 (2011)
26. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
27. Xia, Y., Tang, N., Hussain, A., Cambria, E.: Discriminative bi-term topic model for headline-based social news clustering. In: The Twenty-Eighth International Flairs Conference (2015)
28. Xiao, H., Wang, X., Du, C.: Injecting structured data to generative topic model in enterprise settings. Advances in Machine Learning pp. 382–395 (2009)
29. Zhu, J., Ahmed, A., Xing, E.P.: Medlda: maximum margin supervised topic models. The Journal of Machine Learning Research 13(1), 2237–2278 (2012)