Regularizing Flat Latent Variables with Hierarchical Structures

Rongcheng Lin[+] **, Huayu Li**[+] **, Xiaojun Quan**[†] **, Richang Hong**[⋆] **, Zhiang Wu**[∓] **, Yong Ge**[+]
[+]UNC Charlotte. Email: {rlin4, hli38, yong.ge}@uncc.edu,
[†] Institute for Infocomm Research. Email: quanx@i2r.a-star.edu.sg
[⋆] Hefei University of Technology. Email: hongrc@hfut.edu.cn
[∓] Nanjing University of Finance and Economics. Email: zawu@seu.edu.cn

Abstract

In this paper, we propose a stratified topic model (STM). Instead of directly modeling and inferring flat topics or hierarchically structured topics, we use the stratified relationships in topic hierarchies to regularize the flat topics. The topic structures are captured by a hierarchical clustering method and play as constraints during the learning process. We propose two theoretically sound and practical inference methods to solve the model. Experimental results with two real world data sets and various evaluation metrics demonstrate the effectiveness of the proposed model.

1 Introduction

Probabilistic topic models, such as PLSA [Hofmann, 1999] and LDA [Blei *et al.*, 2003b], have been widely used to automatically uncover the latent topics in unlabeled data sets, especially in text corpora. In general, a topic model is to describe documents as mixtures over a small number of "flat" hidden topics, each of which is independent and represented as a distribution over words. But the "flat" topics tend to be fused or junky, especially when the topic number becomes large [Chuang *et al.*, 2013]. To address the relationships among topics, in the literature of topic modeling, a lot of attempts have been made to incorporate the relationship information into topic modeling process [Andrzejewski *et al.*, 2009; Blei and Lafferty, 2006; Li and McCallum, 2006; Blei *et al.*, 2003a].

One way is to model and infer the pairwise relationships among topics [Blei and Lafferty, 2006; He *et al.*, 2012]. For instance, CTM [Blei and Lafferty, 2006] uses logistic normal distribution with a covariance matrix to model the pairwise relationships between topics. Similarly SLFA [He *et al.*, 2012] tries to discover the pairwise relationships between latent topics through Sparse Gaussian Graphical Model's precision matrix. Another way is to directly model and infer the optimal hierarchically structured topics, such as hierarchical LDA (hLDA) [Blei *et al.*, 2003a] and pachinko allocation model (PAM) [Li and McCallum, 2006]. In hLDA, topics are organized over a tree. Each document is assigned a path through the tree, and each word of the document is sampled from a mixture over topics in the path. PAM uses directed acyclic graph (DAG) to represent topic hierarchies. Each node in the inner layer is a distribution over all nodes on the next level and each of nodes in the leaf layer is a distribution over words.

In this paper, instead of directly modeling the pairwise relationships or hierarchical structures among topics, we use the stratified relationships in topics to regularize the flat topics after capturing the topic structures with a hierarchical clustering method. The hierarchical structure of topics does represent more natural relationships among topics than the pairwise one. By incorporating the stratified relationships that exist among topics into the modeling process, we would be able to uncover more cohesive and meaningful topics. To this end, we develop a stratified topic model (STM), where the hierarchical structure of topics is captured by a tree. Each layer of the tree forms a mixture over different numbers of topics to generate all tokens in a corpus and hierarchical relationships play as constraints during the overall learning process. Along this line, two theoretically sound and practical inference methods are proposed to solve the conceptually nonconjugate model and a heuristic algorithm is developed to construct the tree structure of topics. Finally, we conduct extensive experiments with two real-world data sets. Results based on different evaluation metrics demonstrate that our model could outperform classical topic model methods.

2 Related Work

Traditional probabilistic topic models such as PLSA [Hofmann, 1999] and LDA [Blei *et al.*, 2003b] aim to automatically reduce unlabeled data sets into linear combinations of "flat" latent topics, which are essentially independent from each other. These models have been widely used in many research and applied domains such as natural language processing, information retrieval and computer vision [Foulds and Smyth, 2013; Tang *et al.*, 2013; Gao *et al.*, 2011; Berg *et al.*, 2004]. Moreover, many more extensions and adjustments have been proposed to take into account external information or aspects for better modeling different data sets, such as author-topic model [Rosen-Zvi *et al.*, 2004], DiscLDA [Lacoste-Julien *et al.*, 2008], MG-LDA [Titov and McDonald, 2008], et al. For instance, in addition to the relationships among documents, topics and words, author-topic model [Rosen-Zvi *et al.*, 2004] explores the relationship between authors and them as well.

Many works have attempted to directly model the pairwise relationships among topics, such as CTM[Blei and Lafferty, 2006], SLFA[He *et al.*, 2012]. By capturing the correlation between two topics, this type of models usually lead to better fitting on many data sets. Meanwhile, hierarchical topic modeling techniques aim to directly uncover hierarchical structures that exist in topics. These techniques such as hLDA[Blei *et al.*, 2003a], PAM [Li and McCallum, 2006], HDP [Teh *et al.*, 2006], and hPAM [Mimno *et al.*, 2007] focus on inferring the optimal hierarchical topic structures. Different from all the above approaches, our new model STM uses a hierarchical clustering method to capture hierarchial structures of topics and leverage the stratified relationships among original "flat" topics to regularize flat variables during the learning process.

3 Stratified Topic Model

Our stratified topic model (STM) essentially consists of three-step modeling. First, we train the initial model by assuming no hierarchical structure exists and learn a group of topics. Then a heuristic hierarchical clustering method is applied to construct a stratified topic tree based on topic-word samples. Finally, we refit data by regularizing the initial topics with the built topic tree and learn more cohesive and meaningful topics.

3.1 Notations and Definitions

Given a corpus with M documents and V unique words, probabilistic topic modeling techniques aim to uncover the K latent topics, each of which is represented as a distribution ϕ_k over all words. Each document d_i with N_i words is associated with a latent topic distribution θ_i, which often has a prior distribution with parameters α in Bayesian models. And each word w_{ij} is considered to be drawn from a mixture over all latent topics.

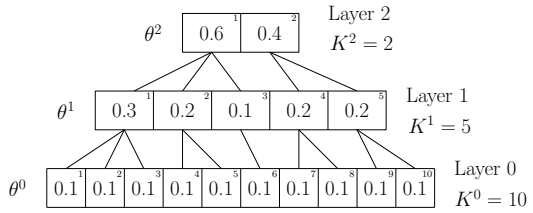

Figure 1: An example of three-level STT.

In STM, the flat topics are organized into a $\mathcal{L}_T$-level stratified topic tree (STT) $\mathcal{T}$, which can be defined as a rose (non-binary) tree associated with a set of tree-consistent partitions over the observed words [Blundell *et al.*, 2010]. Different from the tree structures in hLDA and PAM, each layer l of $\mathcal{T}$ forms a mixture over the K^l latent topics to generate all the words in the document d_i. And this stratified topic tree defines the hierarchical relationships among topic distributions at different layers. The topic proportion of topic k at layer l, $\theta^l[k]$ equals to the sum over the topic proportions of all its leaves. Figure 1 shows an example of the hierarchical relationships of topic distributions defined by a three-level STT.

To describe the hierarchical structure, we use $Par(l,k)$ to denote the parent topic at layer l for leaf topic k and $Leaf(l,k)$ to denote the leaf topic set for topic k at layer l. For example, in the three-level STT shown at Figure 1, $Par(1,5) = 2$ and $Leaf(2,2) = \{7,8,9,10\}$.

3.2 Graphical Model and Generative Process

With the fixed hidden Stratified Topic Tree $\mathcal{T}$ and the initial model hyper-parameters, α and β, as shown in Figure 2, the Stratified Topic Model try to maximize the joint generating likelihood of all the words in document d_i in different tree layers. In other words, each word is sampled $\mathcal{L}_T$ times.

$$P(\mathbf{w}_i^{(0)}, ..., \mathbf{w}_i^{(\mathcal{L}_T-1)} | \alpha, \beta, \mathcal{T})$$

The dependency and model assumptions are illustrated in following generative process:

A. For each tree layer $l \in \{0, ..., \mathcal{L}_T - 1\}$
  a. For each topic $k \in \{1, ..., K^l\}$
  draw the topic-word distribution:$\phi_k^l \sim Dir(.|\beta^l)$
B. For each document $i \in \{1, ..., M\}$
  draw the topic distribution: $\theta_i \sim Dir(.|\alpha)$
  b. For each tree layer $l \in \{0, ..., \mathcal{L}_T - 1\}$
  - For each position $j \in \{1, ..., N_i\}$
  draw a topic assignment: $z_{ij}^l \sim Mult(\theta_i^l)$
  draw a word: $w_{ij} \sim Mult(\phi_{z_{ij}^l}^l)$

In the above process, $\theta_i^l[k] = \sum_{t \in Leaf(l,k)} \theta_i[t]$.

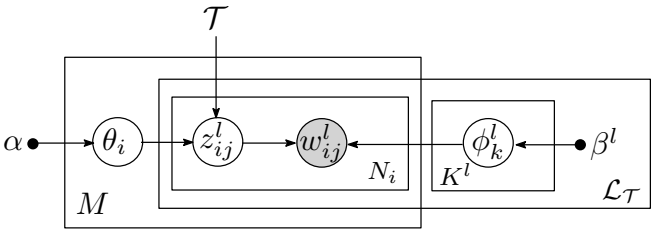

Figure 2: Graphical representation of STM.

We would like to emphasize that the ϕ^{high} at high layer is similar but still different than the combination of the ϕ^{low} at low level due to the constraints of Stratified Topic Tree. Each layer of the model forms a separate "perspective" of the observed words. By integrating the information from different perspectives, words can be better explained and predicted.

3.3 Inference Methods

The stratified topic model is not a conventional conjugate probabilistic model. The posterior distribution of θ_i conditional on the stratified samples is not tractable due to the no-close-form integration over sum of multi variables. In the following paragraphs, we will introduce two methods to infer the parameters of our model: EM for MAP and Relaxed Variational Inference Approximation.

EM for MAP

In the standard EM framework, the object function is maximized iteratively with respect to hidden variables and model parameters. As for STM, in E step, the posterior probability of hidden variable z_{ij} for each term w_{ij} in layer l can be calculated as:

$$p(z_{ij}^l|\theta_i, \phi^l, w_{ij}) \propto p(z_{ij}^l|\theta_i)p(w_{ij}|\phi_{z_{ij}^l}^l) \quad (1)$$

In M step, the parameters (ϕ, θ) are updated using the MAP estimation $(\hat{\phi}_{map}, \hat{\theta}_{map})$. Specifically, We define

$$n_{ikv}^l = \sum_{j=1}^{N_i} I(w_{ij} = v)p(z_{ij}^l = k) \quad (2)$$

where I(.) is the indicator function. Also, for simplicity, we use the dot to denote sum over an index, e.g. $n_{..v}^l = \sum_i^M \sum_k^{K^l} n_{ikv}^l$. Then the MAP estimation of ϕ_k^l is:

$$\hat{\phi}_{map,k}^l[v] = \frac{n_{.kv}^l + \beta_v^l}{n_{.k.}^l + \sum_v \beta_v^l} \quad (3)$$

To formulize the MAP estimation of θ_i, we define a statistic:

$$\mathcal{N}_{ik}^{(l)} := \begin{cases} n_{ik}^0 + \alpha_k & \text{if } l = 0 \\ \mathcal{N}_{ik}^{(l-1)} + \tau_{ik}^l * n_{i\{Par(l,k)\}.}^l & \text{if } l > 0 \end{cases} \quad (4)$$

where $k \in \{1, ..., K^0\}$. And τ_{ik}^l is defined as:

$$\tau_{ik}^l := \frac{\mathcal{N}_{ik}^{(l-1)}}{\sum_{t \in Leaf(l,Par(l,k))} \mathcal{N}_{it}^{(l-1)}} \quad (5)$$

which is the coefficient of allocating the statistic $n_{i\{Par(l,k)\}.}^l$ at layer l to the k_{th} leaf layer node.

We show in Appendix A that the MAP estimation of θ_i is:

$$\hat{\theta}_{map,ik} \propto \mathcal{N}_{ik}^{(\mathcal{L}_T - 1)} \quad (6)$$

As a classic probabilistic topic model, PLSA is considered to be the MAP-estimated LDA model under a uniform dirichlet prior [Girolami and Kabán, 2003]. Similarly, it is easy to find that our STM with MAP estimation is exactly the stratified extension of PLSA.

Relaxed Variational Inference Approximation

The key idea underlying variational inference is to approximate the posterior distributions of hidden variables using a family of distributions with free parameters via minimizing the Kullback-Leibler divergence[Jordan *et al.*, 1999]. In this paper, as shown in Figure 3, we introduce the factorized variational distribution for hidden variables $(\theta_i, \mathbf{z}_i^l)$ in each document d_i as

$$q(\theta_i, \mathbf{z}_i^l|\gamma_i, \eta_i^l) = q(\theta_i|\gamma_i) \prod_l \prod_j q(z_{ij}^l|\eta_{ij}^l) \quad (7)$$

where θ_i subjects to K^0-dimensional Dirichlet distribution $Dir(.|\gamma_i)$ and z_{ij}^l subjects to K^l-dimensional multinomial distribution $Mult(\eta_{ij}^l)$.

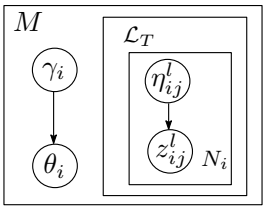

Figure 3: Graphical representation of the variational distribution to approximate the posterior in STM.

In variational inference, minizing the KL divergence between the variational posterior distribution and the true posterior distribution is equivalent to maxmizing the lower bound with respect to free parameters. By applying the Jensen's inequality, the log-likelihood is bounded as

$$\log p(\mathbf{W}_i|\alpha, \beta) \geq E_q(\log p(\mathbf{W}_i, \theta_i, \mathbf{z}_i|\alpha, \beta)) - E_q(\log q(\theta_i, \mathbf{z}_i))$$

In variational EM approach, The lower bound is maximized iteratively with respect to variational parameters $\{\gamma, \eta\}$ (E step) and model parameters $\{\alpha, \beta\}$ (M step). In E step, similar to derivation in [Blei *et al.*, 2003b], we obtain the updating equation for η_{ij}^l:

$$\eta_{ijk}^l \propto \phi_{kw_{ij}}^l exp(\Psi(\tilde{\gamma}_{ik}^l)) \quad (8)$$

By approximating the trigamma function by $\Psi'(x) \approx \frac{c}{x}$, the updating rule for γ_{ik} is:

$$\gamma_{ik} = \mathcal{N}_{ik}^{(\mathcal{L}_T - 1)} \quad (9)$$

where $\mathcal{N}_{ik}^{(\mathcal{L}_T - 1)}$ is the statistic defined at Equation (4). In other words, the posterior distribution is approximated by the Dirichlet distribution whose expectation equals to the MAP estimation. In M step, the model parameters $\{\alpha, \beta\}$ have exactly the same updating rules as those of LDA.

3.4 Stratified Topic Tree

Instead of inferring the tree structure, parameters and hidden variables simultaneously, in STM, the hierarchical topic structure is obtained separately after the initial model is learned. Then the problem to construct a proper topic tree can be regarded as a hierarchical clustering problem over a group of topic-word samples. Although many existing hierarchical clustering techniques can be applied, in this paper, we propose to construct the stratified topic tree using Common Alpha Similarity.

Considering a set of samples of word v over K topics $\{n_{.1v}, n_{.2v}, ..., n_{.Kv}\}$, we can define the generative process for these samples. Samples come from a K-dimensional multinomial distribution parameterized on θ and θ comes from Dirichlet prior with parameter common α. Then the joint probability of these samples is:

$$P(\mathbf{n}_v) = \frac{N_v!}{\prod_k n_{.kv}!} \frac{\Gamma(K\alpha) \prod_k \Gamma(n_{.kv} + \alpha)}{\Gamma(N_v + K\alpha) \prod_k \Gamma(\alpha)} \quad (10)$$

where $N_v = \sum_k n_{.kv}$.

By assuming the samples of the words are independent, the joint probability of all the samples is the product of probability of per-word samples. Then to measure the similarity between topics, we use the Common Alpha Similarity defined

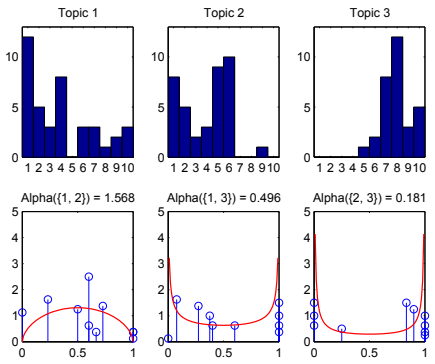

Figure 4: An example of the Common Alpha Similarity between each pair of the above 3 topics with 10 words.

as:

$$\hat{\alpha} = \arg\max_{\alpha} \prod_{v} P(\mathbf{n}_v) \tag{11}$$

It is not hard to find that topics with more similar samples would result in a larger common alpha (as shown in Figure 4). To figure the maximum-likelihood estimation of parameters in Dirichlet-multinomial (Polya) distribution, [Minka, 2000] came up with a fix point iteration for the K-dimensional parameter. As for the common alpha, by using the same bounds, we can derive the similar fix point iteration:

$$\alpha^{new} = \alpha \frac{\sum_v \sum_k (\Psi(n_{.kv} + \alpha) - \Psi(\alpha))}{\sum_v K(\Psi(N_v + K\alpha) - \Psi(K\alpha))} \tag{12}$$

where $\Psi(.)$ is the digamma function. We use $Alpha(\{k_1, ..., k_n\})$ to denote the Common Alpha Similarity of topic set $\{k_1, ..., k_n\}$.

We define a ratio and use it as a criterion for our hierarchical clustering as:

$$\lambda = \frac{Alpha(\{Leaves(T_i) \cup Leaves(T_j)\})}{Alpha(\{Roots(T)\})}. \tag{13}$$

To construct a stratified topic tree, we set a series of ratio λ expectations, one for each layer of the tree. Then following the typical hierarchical clustering scheme, we merge two clusters with the largest Common Alpha Similarity score $Alpha(\{Leaves(T_i) \cup Leaves(T_j)\})$ into a new cluster until the ratio λ is smaller than the expected value for each layer.

3.5 Algorithm

As discussed before, the algorithm consists of 3 steps of modeling:

1. Training the initial STM by assuming that no hierarchical structure exists. Evidently, STM actually degrades into LDA based on this assumption.
2. Building the Stratified Topic Tree using the topic-word samples obtained in the first step. Although we use the Common Alpha Similarity in this paper, a variety of similarity metrics can be utilized to explore the hierarchical structure, which make our STM extensible and flexible in different scenarios.
3. Retraining STM with the constructed stratified topic tree in the second step. Since we build the STT based on topic-word samples, to make sure the consistent model is built, we use the topic-word samples to estimate the initial model parameters $\{\phi^l\}$ (Equation 14) before the EM optimization iterations as:

$$\phi_k^l[v] = \frac{\sum_{t \in Leaf(l,k)} n_{.tv} + \beta_v^l}{\sum_{t \in Leaf(l,k)} n_{.t.} + \sum_v \beta_v^l} \tag{14}$$

4 Experimental Results

4.1 Experimental Setup

Two real-world data sets are used to evaluate the performance of different algorithms: 20 Newsgroups[1] and Encyclopedia Articles[2]. The 20 Newsgroups data set has be manually labelled in a hierarchical way. All documents are almost evenly partitioned into 20 different groups, each corresponding to a different topic. These 20 different groups have been further organized into 6 different larger clusters. Thus there is a two-layer hierarchy among all documents. The Encyclopedia Articles include a small subset of Grolier encyclopedia articles. There are approximately $3,1000$ articles. But there is no labelling information available for this data set.

Typical preprocessing steps are taken for both data sets: *1)* remove punctuations, stopwords and words that have occurred less than 10 times in the whole corpus; *2)* stem the words; *3)* remove those documents containing less than 10 terms; *4)* 40% of documents are selected and used as the testing set. As for 20 Newsgroups data set, it has been already partitioned into training set (60% of documents) and testing set (40% of documents) based on the time. For Encyclopedia Articles data set, we randomly select 40% of documents for testing and use the rest for training. Some statistical characteristics of these data sets are summarized in Table 1.

Name	W	D_{train}	N_{train}	D_{test}
News	15,546	11,269	1,206,449	7,505
Articles	9,557	16,594	1,962,392	11,062

Table 1: Statistics of data sets

While STM explores and leverages hierarchical structures, the outputs of STM are still "flat" yet regularized topics. Thus, in our experiments, we compare STM model with LDA via MAP estimation (PLSA), and LDA via VB estimation. We also compare STM with CTM and HDP[3], which have the same output as STM. As for MAP algorithms, including LDA_{MAP} and STM_{MAP}, the hyper parameters are fixed as $\alpha = 1.1$ and $\beta = 1.01$ as many previous works did. And we set 200 EM iterations with random initialization for each model. As for HDP, we set up 10000 iterations to train the model and extra 1000 iterations to estimate the topic distribution. As for Variational algorithms, including LDA_{VB},

[1] http://qwone.com/~jason/20Newsgroups/

[2] http://www.cs.nyu.edu/~roweis/data.html

[3] We use the available implementation at http://www.cs.columbia.edu/~blei/topicmodeling.html

STM_{VB} and CTM, all models are fitted using initial hyper parameters $\alpha = 0.1$ and $\beta = 0.01$ with random initialization. And we exhaustively run 100 variational EM iterations or until the relative change of lower bound being less than 10^{-5} to ensure the convergence. We find that 2-level STM is sufficient enough to model both data sets, thus in the following experiments the number of layers of stratified topic tree is fixed as $\mathcal{L}_T = 2$. Finally, the stratified topic tree is built by empirically setting the parameter λ as $\lambda = 1.3$.

4.2 Predictive Perplexity

Perplexity, which is equivalent to the inverse of the geometric mean per-word likelihood in algebra, is a conventional metric for evaluating the performance of topic models. Different calculation strategies have been proposed [Wallach *et al.*, 2009]. In this paper, we follow the testing framework of [Wang and Blei, 2013], [Asuncion *et al.*, 2009] and [Blei and Lafferty, 2007], which splits each held-out document into two halves $(\mathbf{w_1}, \mathbf{w_2})$, estimates the document topic distribution with the first half and measures the predictive perplexity using the second half. Also words in each document are randomly shuffled before splitting to get rid of the sequential dependence among tokens. The predictive perplexity is measured using the second half of words as

$$Per(D) = exp(-\frac{\sum_i \sum_{w \in \mathbf{w}_{i2}} \log p(w|\mathbf{w}_{i1})}{\sum_i N_{i2}}), \qquad (15)$$

where N_{i2} is the word count of the second half of document.

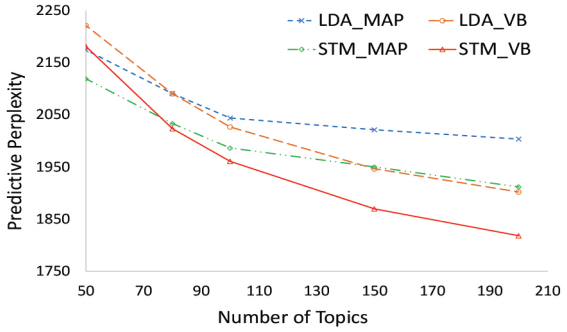

Figure 5: Predictive Perplexity on the 20 Newsgroups

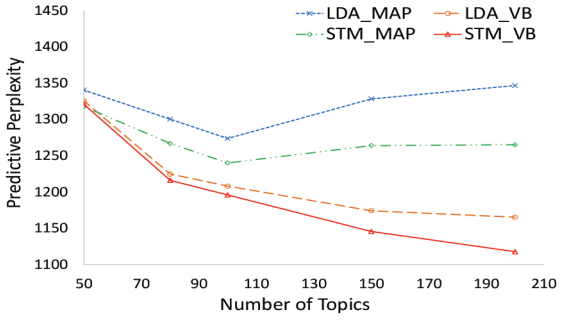

Figure 6: Predictive perplexity on the Encyclopedia Articles

In Figure 5 and Figure 6, we show the comparisons of predictive perplexity among four algorithms with 20 Newsgroups data set and Encyclopedia data set. As can be seen, STM via MAP or Variational Inference could outperform LDA via MAP or variational inference over different numbers of topics. In other words, STM could better predict the remaining words in documents than LDA, which empirically testifies that using stratified relationships to regularize flat topics is effective for modeling a corpus. In fact, we also got the results of CTM on both data sets. It turns out that CTM consistently underperforms LDA with respect to the predictive perplexity. For instance, the predictive perplexity of CTM is 2274.03 when the topic number is 100 on 20 Newsgroups data set. In addition, as shown in Figure 6 and Figure 6, STM with variational inference leads to better performance than STM with MAP because MAP estimation may lead to the typical overfitting. And we can observe the similar comparison trend between two inference methods of LDA from Figures 5 and 6.

4.3 Classification

As a way of dimension reduction, each document could be represented by a lower-dimensional feature vector $p(\theta_i|\mathbf{w_i})$ after topic modelling. An interesting and practical task is to conduct classification based on the learned topics. Specifically, after we train all topic models with the training set of 20 Newsgroups data, the trained models are used to infer the topic distribution of documents in both training and testing set. To make the comparison fair, in STM, we only use the inferred topic distributions at the leaf layer to train a classifier. Thus all topic models transfer each document to a feature vector (i.e., a distribution over topics) with the same dimension. We train two widely used classifier, i.e., logistic regression and SVM in the software LIBLINEAR[4], with the lower-dimensional feature vectors of training set and evaluate the classification accuracy with the lower-dimensional feature vectors of testing set.

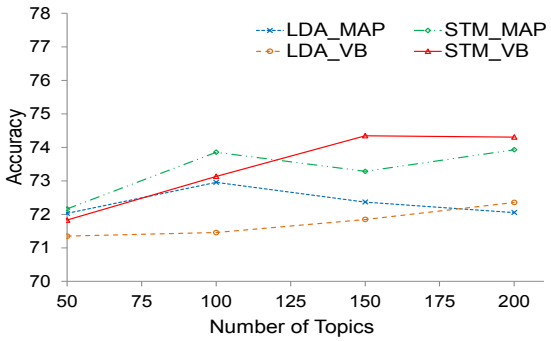

Figure 7: Logistic Regression Classification Performance on the 20 Newsgroups

As shown in Figure 7, we can see that STM could outperform LDA for classification, especially when the topic number becomes larger. The results suggest that STM can better

[4]http://www.csie.ntu.edu.tw/∼cjlin/liblinear/

preserve the latent feature information for classification than LDA. We also conduct the classification task with CTM and HDP in a similar way. But the classification performance via CTM and HDP is worse than that of LDA on 20 Newsgroups data. For instance, CTM leads to 69.40% accuracy on the testing set when the topic number is 100 and HDP leads to only 67.59% with 174 topics. In addition, we observe similar comparison trends among these algorithms with SVM classifier, but we omit the results due to the limit of space.

4.4 Clustering

As we described in the classification task, each document in the training and testing sets is represented by a lower-dimensional feature vector with the same dimension via individual topic model. Since we do not really need a testing set for clustering task, we merge the training and testing sets together and then conduct clustering over the whole data set. Particularly we use the CLUTO[5] to cluster the merged set into 20 clusters. Figure 8 shows the clustering performance using vcluster in CLUTO with cosine similarity and partitional clustering algorithm. From the results we can find that compared to other models, STM leads to better purity score, which means STM can effectively preserve the feature information that is useful for clustering task.

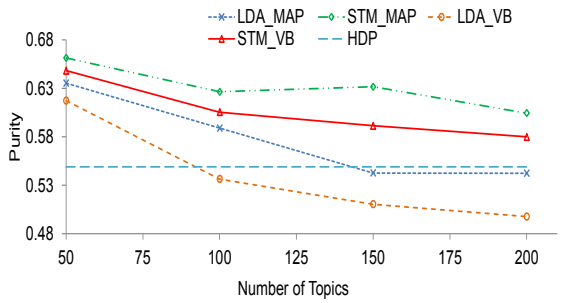

Figure 8: Clustering performance on the 20 Newsgroups

4.5 Additional Experiments

One interesting question about STM is how to set the number of middle-layer topics. In this experiment, we fix the leaf topic number as 100, retrain the 2-level STM with a series of middle topic numbers, and measure the perplexity of each learned model on the 20 Newsgroups data set. From the results shown in Figure 9 we can find that there are generally two local minimums: one locating around 8 and the other one locating around 23. The result is very consistent with the known two-level hierarchical labeling (i.e., 20 small groups and 6 big clusters) of 20 Newsgroups data. The results indicates that our model can effectively capture the useful hierarchical structure information.

5 Conclusions

In this paper, we investigated the problem of leveraging hierarchical structures for better topic modeling. Instead of directly modeling and inferring hierarchical structured topics, we applied a hierarchical clustering method to capture the stratified relationships among flat topics. These relationships were incorporated into the learning process, which is essentially to regularize flat latent variables. We developed two inference methods, i.e., EM for MAP and relaxed variational inference approximation, to solve our model. As shown in our experiments, in addition to better model fitting, STM could also uncover more cohesive and meaningful topics that can help us better explore the semantics of a large corpus and get more robust clustering and classification results.

[5]http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

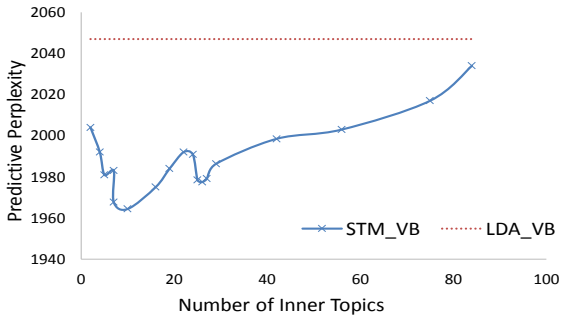

Figure 9: Inner topic number sensibility of 2-level STM with 100 leaf topics on 20 Newsgroups

Acknowledgements

This research was supported in part by National Institutes of Health under Grant 1R21AA023975-01 and National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035.

A MAP Estimation of θ_i

$$\hat{\theta}_i = \arg\max_{\theta_i} \sum_{k=1}^{K^0} \alpha_k \log \theta_{ik} + \sum_{l=0}^{\mathcal{L}_T-1} \sum_{k=1}^{K^l} n^l_{ik.} \log \theta^l_{ik} \quad (16)$$

with the constraint $\sum_{k=1}^{K^0} \theta_{ik} = 1$. To solve the problem, we add the Lagrange multipliers. Then the gradient with respect to θ_{ik} is:

$$\frac{\tilde{\mathcal{L}}[\theta_i]}{\theta_{ik}} = \frac{\alpha_k + n^0_{ik.}}{\theta_{ik}} + \sum_{l=1}^{\mathcal{L}_T-1} \frac{n^l_{i\{Par(l,k)\}.}}{\theta^l_{i\{Par(l,k)\}}} + \lambda \quad (17)$$

Set the derivatives to zero and we get a group of equations. Considering that $\{k_1, ..., k_n\}$ share the same parent node at layer 1, we can derive proportion equation:

$$\frac{\mathcal{N}^0_{ik_s}}{\sum_t^n \mathcal{N}^0_{ik_t}} = \frac{\theta_{ik_s}}{\sum_t^n \theta_{ik_t}} \quad \forall s \in \{1, ..., n\} \quad (18)$$

where $\mathcal{N}^l_{ik}$ is defined at Equation 4. Then the equations can be simplified as:

$$\frac{\tilde{\mathcal{L}}[\theta_i]}{\theta_{ik}} = \frac{\mathcal{N}^1_{ik}}{\theta_{ik}} + \sum_{l=2}^{\mathcal{L}_T-1} \frac{n^l_{i\{Par(l,k)\}.}}{\theta^l_{i\{Par(l,k)\}}} + \lambda = 0 \quad (19)$$

By recursively applying this process, we finally get the MAP estimation for θ_i: $\hat{\theta}_{ik} \propto \mathcal{N}^{\mathcal{L}_T-1}_{ik}$

References

[Andrzejewski *et al.*, 2009] David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32, 2009.

[Asuncion *et al.*, 2009] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 27–34, 2009.

[Berg *et al.*, 2004] Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee Whye Teh, Erik G. Learned-Miller, and David A. Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 848–854, 2004.

[Blei and Lafferty, 2006] David M. Blei and John D. Lafferty. Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

[Blei and Lafferty, 2007] David M. Blei and John D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35, 2007.

[Blei *et al.*, 2003a] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, pages 17–24, 2003.

[Blei *et al.*, 2003b] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Blundell *et al.*, 2010] C. Blundell, Y. W. Teh, and K. A. Heller. Bayesian rose trees. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, pages 65–72, 2010.

[Chuang *et al.*, 2013] Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on Machine Learning*, pages 612–620, 2013.

[Foulds and Smyth, 2013] James R. Foulds and Padhraic Smyth. Modeling scientific impact with topical influence regression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 113–123, 2013.

[Gao *et al.*, 2011] Jianfeng Gao, Kristina Toutanova, and Wen tau Yih. Clickthrough-based latent semantic models for web search. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 675–684, 2011.

[Girolami and Kabán, 2003] Mark Girolami and Ata Kabán. On an equivalence between plsi and lda. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 433–434, 2003.

[He *et al.*, 2012] Yunlong He, Yanjun Qi, Koray Kavukcuoglu, and Haesun Park. Learning the dependency structure of latent factors. In *Advances in Neural Information Processing Systems*, pages 2375–2383, 2012.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[Jordan *et al.*, 1999] Michael I. Jordan, Zoubin Ghahramani, and et al. An introduction to variational methods for graphical models. In *Machine Learning*, pages 183–233, 1999.

[Lacoste-Julien *et al.*, 2008] Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, pages 897–904, 2008.

[Li and McCallum, 2006] Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 577–584, 2006.

[Mimno *et al.*, 2007] David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 633–640, 2007.

[Minka, 2000] Thomas P. Minka. Estimating a dirichlet distribution. Technical report, 2000.

[Rosen-Zvi *et al.*, 2004] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

[Tang *et al.*, 2013] Jian Tang, Ming Zhang, and Qiaozhu Mei. One theme in all views: Modeling consensus topics in multiple contexts. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 5–13, 2013.

[Teh *et al.*, 2006] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[Titov and McDonald, 2008] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, pages 111–120, 2008.

[Wallach *et al.*, 2009] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112, 2009.

[Wang and Blei, 2013] Chong Wang and David M. Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(1):1005–1031, 2013.