Topic Modeling with Nonparametric Markov Tree

Haojun Chen HAOJUN.CHEN@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

David B. Dunson DUNSON@STAT.DUKE.EDU
Department of Statistical Science, Duke University, Durham, NC 27708, USA

Lawrence Carin LCARIN@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

Abstract

A new hierarchical tree-based topic model is developed, based on nonparametric Bayesian techniques. The model has two unique attributes: (*i*) a child node in the tree may have more than one parent, with the goal of eliminating redundant sub-topics deep in the tree; and (*ii*) parsimonious sub-topics are manifested, by removing redundant usage of words at multiple scales. The depth and width of the tree are unbounded within the prior, with a retrospective sampler employed to adaptively infer the appropriate tree size based upon the corpus under study. Excellent quantitative results are manifested on five standard data sets, and the inferred tree structure is also found to be highly interpretable.

1. Introduction

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is widely used to infer low-dimensional latent semantic information (topics) associated with a corpus of documents, where a topic is a distribution over words. While LDA constitutes a powerful modeling paradigm, and has served as the motivation for many subsequent models, a limitation of LDA and many such models is that within the prior no statistical dependencies are assumed between topics (each topic is drawn i.i.d. from a Dirichlet distribution). However, such statistical dependencies typically exist, for example a topic related to soccer and another topic related to football share many words, but these topics are not exactly the same; it is desirable to impose within the model this prior expectation of statistical dependencies between topics. To address this challenge, there has been recent interest in *hierarchical* topic models (Adams et al., 2010; Blei et al., 2010; Chambers et al., 2010; Griffiths et al., 2007; Jenatton et al., 2010; Li & McCallum, 2006). In such settings one may view the model as inferring "meta-topics" that are constituted by integrating modular components ("sub-topics") within a hierarchy. For example, assume that ϕ_s is a probability vector representing a sub-topic, and $\{\phi_s\}$ represents a finite set of such probability vectors; any convex combination of the $\{\phi_s\}$ may be viewed as constituting a meta-topic. Hierarchical models are typically based on modular elements like $\{\phi_s\}$, the inter-relationships between which are often constituted in a tree-based manner. In the context of the soccer/football illustration above, each may be a meta-topic, constituted by a convex combination of sub-topics from $\{\phi_s\}$; these two meta-topics will likely share some components $\{\phi_s\}$, but not all.

An important advantage of using such hierarchical and modular models is that different meta-topics (which may be defined, for example, by a branch of a tree) share components of the set $\{\phi_s\}$, and therefore there is a significant opportunity to borrow statistical strength efficiently across a corpus. While two meta topics may be distinct, they may share components of $\{\phi_s\}$, and therefore the available data are shared to a desirable extent when learning $\{\phi_s\}$.

In the nested Chinese restaurant process (nCRP) topic model (Blei et al., 2010), each node in the tree is characterized by a sub-topic, and a document is generated from one path (branch) through the tree, from the root to a leaf. In the tree-structured stick breaking process (TSSB) (Adams et al., 2010), each node is a unique dis-

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

tribution over topics and a document is generated from one node of the tree. These hierarchical topic models yield good performance. However, they sometimes may be too rigid. For example, in the nCRP model (Blei et al., 2010) there is a single root node, and children nodes may only have a single parent. This means that all descendent sub-topics from parent $p1$ must be distinct from the descendants of parent $p2$, if $p1 \neq p2$. Some of these distinct sets of children from different parents may be redundant, and this redundancy can be removed if a child can have more than one parent; this is one motivation of our proposed model. To get a sense of our model, and to observe its distinction from tree-based models such as that in (Blei et al., 2010), consider Figure 1. This figure depicts a small subset of the sub-topics inferred at two adjacent scales, for a real document corpus, considered in detail when presenting results. A Markov transition process is inferred to move from sub-topics at one scale to those at the next scale, and from Figure 1 it is possible for a sub-topic to have two or more parents, linked to the parents in a statistical (Markov) sense. In Figure 1, the sub-topics at the bottom distinguish between "software" (left) and "hardware" (right); at the top layer the left sub-topic corresponds to general systems/technology, while the right sub-topic corresponds specifically to computers and PCs. By sharing children across multiple parent nodes, there is not a need to have children specialized to a distinct parent, and therefore potentially have nearly duplicate but decoupled children. This model flexibility is anticipated to enhance the sharing of statistical strength, in the sense discussed above. To our knowledge, none of the hierarchical and tree-based models developed previously have this flexibility.

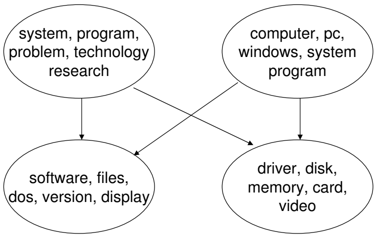
system, program, problem, technology research
computer, pc, windows, system program
software, files, dos, version, display
driver, disk, memory, card, video

Figure 1. Illustrative example topics, inferred form the *20 Newsgroup* document corpus.

Each of the sub-topics reflected in $\{\phi_s\}$ are probability vectors over the vocabulary. In most existing tree-based hierarchical models there is typically nothing within the model prior (Adams et al., 2010; Blei et al., 2010; Griffiths et al., 2007) that places restrictions on these multiple probability vectors; hence, when drawing the ϕ_s from the prior, there may be significant duplication in word usage, in the sense that a particular word may be probable in many of the ϕ_s. This problem may be partially mitigated in the posterior for $\{\phi_s\}$, after analyzing the corpus; however, it is desirable to impose as much structure as possible within the prior, such that there is less reliance on the data to infer anticipated phenomena. So motivated, within the proposed model we constitute a new framework, imposing that if a particular word is present in one or more sub-topics at a particular scale, then this word may not be used for sub-topics at scales deeper in the tree. Among other things, this removal of redundant usage of the same word at multiple scales aids in interpreting the multi-scale sub-topics inferred by the model.

We employ a retrospective sampler (Papaspiliopoulos & Roberts, 2008), within a stick-breaking representation of the Dirichlet process (Ferguson, 1973) and related stick-breaking constructs (Sethuraman, 1994). In this setting the depth and width of the tree is unbounded, and hence the tree structure is inferred nonparametrically from the data. To the authors' knowledge, the form of the retrospective sampler employed in the proposed model is also new to topic modeling.

2. Proposed Model

2.1. Multi-scale Markov tree

We wish to model a corpus composed of D documents; the size of the vocabulary is V. We develop a hierarchical Bayesian model that infers a multi-scale topic construction. The lowest-level scale/resolution is $s = 1$, and we wish more detailed/specific words to be emphasized as one progresses deeper in the tree (*i.e.*, increasing s). Moreover, we wish to impose that if a word is utilized to constitute sub-topics at scale s', then this word is not reused in sub-topics at scales $s > s'$ (*i.e.*, deeper in the tree). The proposed model does not in general have a single root node, as in (Blei et al., 2010), and all children at scale $s+1$ are connected statistically to all parents at scale s, via a Markov process. The depth of the tree, and the width (number of nodes) at each scale are inferred nonparametrically from the data.

For node t at scale s, there is an associated V-dimensional probability vector over words, denoted $\boldsymbol{\phi}_{st}$, this representing a sub-topic. Further, when drawing word i from document d, w_{di}, there is a latent integer $c_{di} \geq 1$ defining which scale w_{di} is drawn from. The generative processes for $\{\boldsymbol{\phi}_{st}\}$ and $\{c_{di}\}$ are discussed below; here we describe the process by which we define the specific node at scale c_{di} from which word w_{di} is drawn.

The probability of utilizing each of the nodes (subtopics) at the first scale $s = 1$ is defined by the document-dependent probability vector $\boldsymbol{\theta}_d$, drawn as

$$\boldsymbol{\theta}_d \sim \mathrm{DP}(\eta, \boldsymbol{\alpha}) \ , \quad \boldsymbol{\alpha} \sim \mathrm{Stick}(\lambda)$$

where $\mathrm{DP}(\eta, \boldsymbol{\alpha})$ represents a Dirichlet process (DP) (Ferguson, 1973) with base measure $\boldsymbol{\alpha}$ and real innovation parameter $\eta > 0$; the expression $\mathrm{Stick}(\lambda)$ represents a stick-breaking process (Sethuraman, 1994) with parameter λ, and the ith component of draw $\boldsymbol{\alpha}$ is $\alpha_i = V_i \prod_{h=1}^{i-1}(1 - V_h)$ with $V_h \sim \mathrm{Beta}(1, \lambda)$. The aforementioned DP draw may be constituted as $\boldsymbol{\theta}_d \sim \sum_{j=1}^{\infty} \pi_j \delta_{\phi_j^*}$, with $\phi_j^* \sim \sum_{k=1}^{\infty} \alpha_k \delta_k$ and with $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ drawn $\boldsymbol{\pi} \sim \mathrm{Stick}(\eta)$. The $\boldsymbol{\theta}_d$ are therefore drawn in a form related to the hierarchical DP (HDP) (Teh et al., 2006), with stick-breaking construction.

The node from which word w_{di} is drawn is manifested via a Markov process. Specifically, the generative process sequentially selects one node at each scale, up to scale c_{di}. When at node m at scale $s \geq 1$, the probability vector $\boldsymbol{p}_m^{(s)}$ defines the probability of which node is transitioned to at scale $s+1$. This probability vector is drawn

$$\boldsymbol{p}_m^{(s)} \sim \mathrm{Stick}(\zeta_s)$$

Drawing $\{\boldsymbol{p}_m^{(s)}\}$ in this manner for all nodes at scale s, a matrix $\mathbf{P}^{(s)}$ is defined, the mth column of which is $\boldsymbol{p}_m^{(s)}$. Note that the matrices $\{\mathbf{P}^{(s)}\}$ for scales $s \geq 1$ are assumed independent of the document d. If the model is truncated to only one scale, this model is essentially the previously developed HDP-based topic model (Teh et al., 2006). The use of additional scales $s > 1$ are meant to capture finer details in the topics, adding flexibility by allowing incorporation of detail to the topics.

We will infer the number of required scales s to represent the corpus, using a retrospective sampler (Papaspiliopoulos & Roberts, 2008), as discussed in Section 2.4. Similarly, a retrospective sampler is also used to draw the $\boldsymbol{p}_m^{(s)} \sim \mathrm{Stick}(\zeta_s)$, and therefore the number of nodes or sub-topics at each scale is also inferred. Hence we infer the depth of the tree, the width of each scale, and a Markovian statistical relationship between all parents at scale s and all children at scale $s + 1$. Gamma hyperpriors are placed on λ and each ζ_s, and hence posterior distributions are also inferred for these parameters.

2.2. Node & scale-dependent word probabilities

Assume that through the aforementioned Markov process to scale c_{di}, node t is arrived at, and it is from this node that word w_{di} is drawn. Hence, word w_{di} is drawn from a multinomial distribution with parameter $\boldsymbol{\phi}_{c_{di}t}$. We now define a generative process for probability vectors $\{\boldsymbol{\phi}_{st}\}$, imposing that if a word has non-zero probability of occurring at scale s', then it has zero probability of occurring at scale $s > s'$.

At each scale $s \geq 1$ we define a V-dimensional binary vector $\boldsymbol{b}_s = (b_{s1}, \dots, b_{sV})^T$. Each of the scale and node dependent probability vectors over words are drawn

$$\boldsymbol{\phi}_{st} \sim \mathrm{Dirichlet}(\gamma \boldsymbol{b}_s)$$

If $b_{sv} = 0$, then word $v \in \{1, \dots, |V|\}$ will have zero probability of being manifested at scale s (for all nodes t); *i.e.*, the vth component of $\{\boldsymbol{\phi}_{st}\}$ will be zero for all t. Therefore, within the model we impose that if the vth component of $\boldsymbol{b}_s$ is non-zero for a particular scale s, then the vth component of $\boldsymbol{b}_{s'}$ is zero for all $s' > s$. Specifically, the generative process is

$$p(b_{sv} = 1 | \boldsymbol{b}_{[s-1]v}) = 1(\boldsymbol{b}_{[s-1]v} = \mathbf{0})\tau_s, \tau_s \sim \mathrm{Beta}(1, \psi_s)$$

where $\boldsymbol{b}_{[s-1]v} = (b_{1v}, \dots, b_{s-1,v})^T$, $1(\cdot)$ is the indicator function, by convention $b_{0v} = 0$ for all $v \in \{1, \dots, V\}$, τ_s is the conditional probability of adding a word to scale s given that it has not been added to any of the previous scales, and $\psi_s \geq 0$ is a hyperparameter controlling the distribution of words at scale s. One may wish to impose a separate prior for each scale-dependent parameter ψ_s, for example favoring smaller τ_s with increasing s (such that the $\{\boldsymbol{\phi}_{st}\}$ are sparser with increasing s, favoring more-detailed words).

A distribution similar to the above $\mathrm{Dirichlet}(\gamma \boldsymbol{b}_s)$ was used for language modeling in sparseTM (Wang & Blei, 2009a) and FTM (Williamson et al., 2010). However, the hierarchical construction specified above, for which words are not reused, is unique to the proposed model.

2.3. Generative process

Figure 2 provides a graphical depiction of the generative process. The generative process of this model can be summarized as follow:

1. For each scale s, for each term v, draw term selector $b_{sv} \sim \mathrm{Bernoulli}(1(\boldsymbol{b}_{[s-1]v} = \mathbf{0})\tau_s), \tau_s \sim \mathrm{Beta}(1, \psi_s)$
2. For each topic $t \in 1, \dots, T_s$ in scale s,
   (a) Draw topic distributions $\boldsymbol{\phi}_{st} \sim \mathrm{Dirichlet}(\gamma \boldsymbol{b}_s)$
   (b) Draw transition matrix $\mathbf{P}^{(s)}$, the t-th column $\boldsymbol{p}_t^{(s)} \sim \mathrm{Stick}(\zeta_s)$

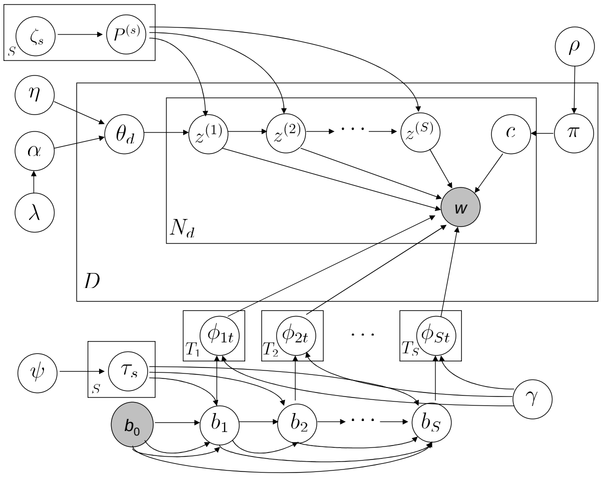

Figure 2. Graphical model representation for multi-scale Markov topic model.

3. Draw stick lengths $\boldsymbol{\alpha} \sim \text{Stick}(\lambda)$, which are the global distribution over topics

4. For document d
   (a) Draw distribution over topics for the first scale $\boldsymbol{\theta}_d \sim \text{DP}(\eta, \boldsymbol{\alpha})$
   (b) For the i-th word:
      i. Draw scale indicator $c_{di} \sim \text{Mult}(\boldsymbol{\pi}_d)$, $\boldsymbol{\pi}_d \sim \text{Stick}(\rho)$
      ii. Draw topic indicator $z_{di}^{(1)} \sim \text{Mult}(\boldsymbol{\theta}_d)$ if $c_{di} \geq 2$, then for $s = 2, \dots, c_{di}$ $z_{di}^{(s)} | (z_{di}^{(s-1)} = m) \sim \text{Mult}(\boldsymbol{p}_m^{(s-1)})$
      iii. Draw word $w_{di} | (c_{di} = s, z_{di}^{(s)} = t) \sim \text{Mult}(\boldsymbol{\phi}_{st})$

A gamma hyperprior is placed on ρ, and $\boldsymbol{\pi}_d$ defines the probability of using each of the scales in the tree, with the probability of scale usage document-dependent. Note that in principle $\boldsymbol{\pi}_d$ is an infinite-dimensional probability vector, and hence the number of scales is unbounded. As discussed in the next subsection, a retrospective sampler (Papaspiliopoulos & Roberts, 2008) is employed, and therefore the model infers the number of scales that are needed for representation of the corpus, just as a similar approach is employed to infer the number of nodes (sub-topics) at each scale.

Note that in our proposed model stick-breaking representations are employed in depth, and also in width, at each scale. In this sense the model is related to the tree-structured stick-breaking model in (Adams et al., 2010). However, in (Adams et al., 2010) an entire document is inferred to reside at one node in the tree, and in this sense the model may be viewed as yielding hierarchical clustering for documents. In the proposed model each node of the model corresponds to a sub-topic, as in (Blei et al., 2010). However, unlike (Blei et al., 2010), children nodes may have multiple parent nodes, in a Markovian statistical sense, yielding a more-flexible construction with more sharing of sub-topics. The model in (Jenatton et al., 2010) has a parent-child tree-based construction similar to that in (Blei et al., 2010), but the finite tree must be specified, or inferred via cross-validation. The imposition, within the prior, of structure on word usage between scales is also unique to the proposed model. Our model is also related to PAM(Li & McCallum, 2006) and GraphLDA(Chambers et al., 2010) in constructing flexible structures for a set of topics learned from text. But our approach differs in that we introduce Markov dependency to construct the paths and the depth and width of the tree are inferred using retrospective sampling.

2.4. Retrospective inference

In this subsection we describe the retrospective sampling scheme to learn S, the depth of the tree, and T_s, the width of the tree at each scale s. Let S be the learned depth of the tree in a given iteration and assume we have already obtain samples of $\{\{\boldsymbol{b}_s\}, \boldsymbol{\rho}, \{\boldsymbol{\phi}_{s,t}\}\}$ for $1 \leq s \leq S$. We learn S by updating each of the c_{di}'s in a Metropolis-Hastings step. When updating c_{di}, the proposed s' is generated from the following distribution

$$q_{di}(s') \propto \begin{cases} \pi_{s'} p(w_{di} | \boldsymbol{\phi}_{s', z_{di}^{(s')}}), & \text{for } s' \leq S \\ \pi_{s'} \mathcal{M}_{di}(S), & \text{for } s' > S \end{cases}$$

where $\mathcal{M}_{di}(S) = \max_{1 \leq s \leq S} \{p(w_{di} | \boldsymbol{\phi}_{s, z_{di}^{(s)}})\}$. The acceptance probability for the proposed s' is $\kappa_{di}(s, s')$, defined as

$$\begin{cases} 1, & \text{if } s' \leq S \text{ and } S' = S \\ \min\{1, \frac{\tilde{c}_{di}(S) \mathcal{M}_{di}(S')}{\tilde{c}_{di}(S') p(w_{di} | \boldsymbol{\phi}_{s, z_{di}^{(s)}})}\}, & \text{if } s' \leq S \text{ and } S' < S \\ \min\{1, \frac{\tilde{c}_{di}(S) p(w_{di} | \boldsymbol{\phi}_{s', z_{di}^{(s')}})}{\tilde{c}_{di}(S') \mathcal{M}_{di}(S)}\}, & \text{if } s' > S \end{cases}$$

where $S' = \max\{\max_{d' \neq d, i' \neq i} \{c_{d',i'}\}, s'\}$ and the normalizing constant $\tilde{c}_{di}(S) = \sum_{s=1}^{S} \pi_s p(w_{di} | \boldsymbol{\phi}_{s, z_{di}}) + \mathcal{M}_{di}(S)(1 - \sum_{s=1}^{S} \pi_s)$. The whole retrospective sampling procedure for c_{di} is summarized in Algorithm 1. As shown in the algorithm, S is learned automatically by calculating the number of unique values of c_{di}'s. A similar retrospective sampling procedure can be developed for learning T_{s+1} by updating $\{\mathbf{P}^{(s)}\}$ (Paisley

& Carin, 2009). The update equations for $\{\mathbf{P}^{(s)}\}$ are provided in the next section.

Algorithm 1 Retrospective Inference for c_{di}

```
for $d = 1$ to $D$ do
  for $i = 1$ to $N_d$ do
    Sample $u_{di} \sim \text{Uniform}(0,1)$ and $s' = 1$
    while $u_{di} > \sum_{l=1}^{S} q_{di}(l)$ do
      $S = S + 1$, sample $\boldsymbol{b}_{S-1}, \{\nu_{dS}\}, \{\boldsymbol{\phi}_{St}\}, z_{di}^{S}$,
      and $\mathbf{P}^{(S-1)}$ from the prior, $\boldsymbol{b}_S = 1 - \sum_{h=1}^{S-1} \boldsymbol{b}_s$,
      and $\pi_{dS} = \nu_{dS} \prod_{h=1}^{S-1}(1 - \nu_{ds})$
    end while
    while $u_{di} > \sum_{l=1}^{s'} q_{di}(l)$ do
      $s' = s' + 1$
    end while
    $c_{di} = s'$ with probability $\kappa_{di}(s, s')$, otherwise,
    leave $c_{di}$ unchanged
  end for
end for
```

3. Model Inference

We provide update equations that are unique for this model. The update equations for the rest of the parameters are similar to those in HDP (Teh et al., 2006).

Define $A_s \triangleq \{v : b_{sv} = 1, v \in \mathcal{V}\}$ to be the set of indices of $\boldsymbol{b}_s$ that are utilized. Let $n_{st}^{(v)}$ denote the number of times that term v has been assigned to topic t in scale s, and let $n_{st}^{(\cdot)}$ denote the number of times that all the terms have been assigned to topic t in scale s.

The Gibbs sampling inference procedure is described as follow:

- Sampling transition matrix $\mathbf{P}^{(s)}$:

$$\begin{aligned} V_{tm}^{(s)} \sim\ & \text{Beta}(1 + \sum_{d,i} 1(z_{di}^{(s+1)} = t, z_{di}^{(s)} = m), \\ & \zeta_s + \sum_{d,i} 1(z_{di}^{(s+1)} > t, z_{di}^{(s)} = m)), \end{aligned}$$

$$p_{1m}^{(s)} = V_{1m}^{(s)}, \quad p_{tm}^{(s)} = V_{tm}^{(s)} \prod_{l=1}^{t-1}(1 - V_{lm}^{(s)})$$

  For new values, set $p_{T_{s+1}+1,m}^{(s)} = 1 - \sum_{l=1}^{T_{s+1}} p_{l,m}^{(s)}$ and draw a new column for $\mathbf{P}^{(s+1)}$ from the prior.

- Sampling Bernoulli parameter τ_s and term scale indicators $\boldsymbol{b}_s$:

$$\begin{aligned} & p(\tau_s, \boldsymbol{b}_s | -) \\ \propto\ & p(\boldsymbol{b}_s | \tau_s, \boldsymbol{b}_{[s-1]}) p(\tau_s | \psi_s) \prod_{(d,i):c_{di}=s} p(w_{di} | \boldsymbol{b}_s) \\ =\ & p(\boldsymbol{b}_s | \tau_s, \boldsymbol{b}_{[s-1]}) p(\tau_s | \psi_s) \int d\boldsymbol{\phi}_{st} \{ p(\boldsymbol{\phi}_{st} | \boldsymbol{b}_s) \\ & \prod_{(d,i):c_{di}=s} p(w_{di} | \boldsymbol{\phi}_{st}) \} \\ =\ & p(\boldsymbol{b}_s | \tau_s, \boldsymbol{b}_{[s-1]}) p(\tau_s | \psi_s) \\ & \prod_{t=1}^{T_s} \frac{\Gamma(\gamma |A_s|) \prod_{v \in A_s} \Gamma(n_{st}^{(v)} + \gamma)}{\Gamma^{|A_s|}(\gamma) \Gamma(n_{st}^{(\cdot)} + \gamma |A_s|)} \end{aligned}$$

  where $|A_s| = \sum_v b_{sv}$ and $\Gamma(\cdot)$ is the gamma function. We can iteratively sample $\boldsymbol{b}_s$ conditioned on τ_s and sample τ_s conditioned on $\boldsymbol{b}_s$ by this joint conditional distribution. In addition, $\boldsymbol{b}_S = 1 - \sum_{h=1}^{S-1} \boldsymbol{b}_s$.

- Sampling topic indicator $z_{di}^{(s)}$:
  For $s = 1$,

$$\begin{aligned} p(z_{di}^{(1)} | -) \propto\ & p(z_{di}^{(1)} | \boldsymbol{\theta}_d) [p(w_{di} | \boldsymbol{\phi}_{1, z_{di}^{(1)}}) 1(c_{di} = 1) \\ & + p(z_{di}^{(2)} | z_{di}^{(1)}) 1(c_{di} > 1)] \end{aligned}$$

  For s ≥ 2,

$$\begin{aligned} p(z_{di}^{(s)} | -) \propto\ & p(z_{di}^{(s)} | z_{di}^{(s-1)}) [p(w_{di} | \boldsymbol{\phi}_{s, z_{di}^{(s)}}) 1(c_{di} = s) \\ & + p(z_{di}^{(s+1)} | z_{di}^{(s)}) 1(c_{di} > s)] \end{aligned}$$

  After normalization, $p(z_{di}^{(s)} | -)$ becomes a multinomial distribution.

4. Empirical Study

In the following experiments, if without other specifications, all the hyperparameters for the gamma distributions are set to Gamma$(10^{-3}, 10^{-3})$ and $\gamma = 1$, with no tuning performed on these parameters. These are the only parameters that need be set in the model, with an approximate posterior distribution estimated for all other parameters, based on the sampler. Within the sampler, we employed 2500 burn-in iterations, and we collected 500 samples after burn-in, taking every fifth sample to approximate the posterior. From a rigorous mathematical perspective, the number of iterations may still be too small to ensure convergence, but in practice we find that they are large enough for achieving reasonable results.

4.1. Quantitative assessment

We compare the performance of our model to LDA (Blei et al., 2003) and nCRP (Blei et al., 2010). We denote the proposed model as HMT, for "hierarchical Markov tree", and we present our model in two forms. The results denoted HMT-SD employ the scale dependency on word usage within sub-topics, as discussed in Section 2.2 (hence HMT-SD is our complete model). The scale-dependent word assumption is optional, and it can be removed if it is undesired. To examine the importance of imposing this scale-dependency to the word usage, we also consider HMT-NSD, in which no scale-dependency is imposed on the use of words within sub-topics; in this case $\boldsymbol{b}_s$ is all ones, for all scales s. The models are examined on the following data sets:

- JCAM: a collection of 536 abstracts from the *Journal of the ACM* from 1987 to 2004 and the vocabulary size is 1539.[1]
- Psy. Review: a collection of 1281 abstracts from *Psychological Review* from 1967 to 2003 and the vocabulary size is 1971.
- 20 Newsgroups: A collection of 3000 documents, randomly selected articles from the *20 Newsgroups* data set; the vocabulary size is restricted to 5957 by a Porter stemmer.[2]
- NIPS: A collection of 1740 *NIPS* articles published from 1988 to 1999, and the vocabulary size is restricted to 7546 by a Porter stemmer.[3]
- Reuters: A collection of 3000 documents, randomly selected documents from the *Reuters-21578* data set, and the vocabulary size is restricted to 1671 by a Porter stemmer.[4]

For all data sets, terms that appear in fewer than six documents were removed.

We first use the test set likelihood to evaluate the predictive performance of HMT. The common method to evaluate predictive performance in topic model is to use cross validation. Here we use five-fold cross validation. To calculate the conditional probability of the test set given the training set, we use the same method and same parameter settings as in nCRP (Blei et al., 2010) and VB-nCRP (Wang & Blei, 2009b). Specifically, we set the depth $S = 3$, $\eta = 1$ and use J samples and compute

$$p(\boldsymbol{w}_1^{\text{test}}, \ldots, \boldsymbol{w}_{D_{\text{test}}}^{\text{test}} | \boldsymbol{w}_1^{\text{train}}, \ldots, \boldsymbol{w}_{D_{\text{train}}}^{\text{train}}, \text{Model}) = \prod_{d,i} \frac{1}{J} \sum_{j=1}^{J} \sum_{s=1}^{S} 1(c_{di}^{(j)} = s) \sum_{t} \hat{\theta}_{dt}^{(j)} \psi_{stw_{di}}^{(j)}$$

where $\hat{\boldsymbol{\theta}}_d^{(j)} = \mathbf{P}^{(s-1)^{(j)}} \ldots \mathbf{P}^{(1)^{(j)}} \boldsymbol{\theta}_d^{(j)}$. Larger log likelihood is better, and the log likelihood is computed by averaging across the collection samples. The mean test set log likelihood values and the standard deviations are shown in Table 1. For the LDA(Blei et al., 2003) results, similar to (Blei et al., 2010), we first run HMT-SD to obtain a posterior distribution over the number of topics, and then run LDA multiple times using the learned range of the number of topics in HMT-SD. The nCRP results are directly from (Wang & Blei, 2009b). In each case, the HMT-SD achieves larger log likelihood than the three alternative models, significantly better than LDA. The performance of HMT-SD is similar to that of nCRP in *JACM* and does better in *Psychological Review.* As stated in nCRP (Blei et al., 2010), for a larger corpora the nCRP may be rigid for constraining to use only a single path, but HMT-SD provides more flexibility in selecting topics. From these results we also note that the imposition of structure on the probability of using words in sub-topics, as a function of scale s yields significant improvements. Specifically, both HMT-NSD and HMT-SD employ the same tree structure, while the former does not impose structure within the prior on word usage in scale-dependent sub-topics.

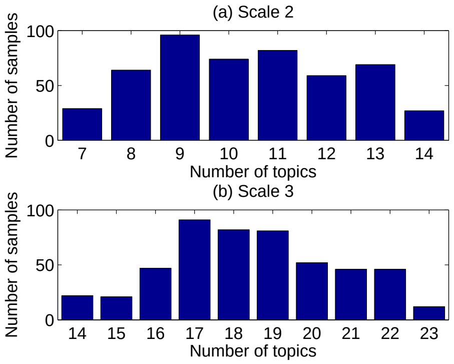

Figure 3. Approximate posterior distribution (histogram) on the number of topics at scales $s = 2$ and $s = 3$, for the proposed model considering the *20 Newsgroup* data.

4.2. Model structure and parsimony

Within the MCMC sampler, each collection sample yields a unique topic tree. To get a sense of the variety of models that are manifested via these collection samples, in Figure 3 we show a histogram of the number of topics at scales $s = 2$ and $s = 3$ for the *20 Newsgroup* data set; similar results were found for the other data sets. Note that these numbers of topics at each scale were inferred efficiently via the retrospective sampler, and a similar distribution is found with respect to the depth of the tree.

[1] http://www.cs.princeton.edu/∼blei/downloads/
[2] http://people.csail.mit.edu/jrennie/20Newsgroups/
[3] http://www.cs.nyu.edu/∼roweis/data.html
[4] http://kdd.ics.uci.edu/databases/reuters21578/

Table 1. Test set per-word log likelihood for the five data sets, and four algorithms. The HMT-SD model is that proposed here, and HMT-NSD is a comparison simplification. The LDA model is from (Blei et al., 2003) and the nCRP model is from (Blei et al., 2010).

	JCAM	Psy. Review	20 Newsgroup	NIPS	Reuters
LDA	-13.6237±0.0159	-14.6483±0.0192	-15.1791±0.0149	-16.2404±0.0183	−13.7142±0.0168
nCRP	-5.3922±0.0052	-5.7834±0.0149	*	*	*
HMT-NSD	-5.6048±0.0059	-5.7530±0.0153	-6.6949±0.0175	-6.8962±0.0178	-5.6575±0.0150
HMT-SD	**-5.2770±0.068**	**-5.4734±0.141**	**-6.2952±0.0135**	**-6.4665±0.0194**	**-5.3746±0.0105**

To further examine the properties of the model, we examine the number of words found in each topic in different scale, on average. The results in Figure 4, for the aforementioned models and data sets, indicate that the HMT-SD model yields a more parsimonious usage of words across topics. In Figure 4 we present mean results, as well as the standard deviation. In each case, the number of terms per topic in HMT-SD is significantly less than that of HMT-NSD. This is attributed to the HMT-SD prior that removes redundancy of words at multiple scales (although there can be redundancy *within* a single scale), and therefore the topics tend to be more focused and less redundant. In addition to aiding the ability to interpret inferred topics, this construction improves quantitative log likelihood results, as discussed above.

4.3. Interpreting the learned tree

In Figure 5 we present an example topic tree inferred from the proposed model. In this analysis a single root node was employed, meant to capture the ubiquitous and non-informative words. The transition probabilities from the root node to the second layer are document-dependent, and all other transition probabilities between layers are document-independent. Within Figure 5, which corresponds to a typical collection sample, we present the top-five most-probable words within each sub-topic, and the arrows represent non-zero transition probabilities. Note that sub-topics at layer $s = 3$ often have multiple parents, in a statistical (Markovian) sense, this representing a unique component of the proposed model.

Careful examination of Figure 5 demonstrates that several interesting relationships are inferred between sub-topics at different scales. For example there appear at layer $s = 3$ to be sub-topics on hockey and baseball, and each of these share a parent at layer $s = 2$ that appears to capture sports in a generic sense. Another interesting example concerns a sub-topic at scale $s = 2$ that focuses on systems and technology, and this is connected to children sub-topics at layer $s = 3$ focusing (separately) on bikes, cars, space travel, electronic devices, software, and computer hardware. Similar trees are manifested by the model in (Blei et al., 2010) (the code for (Blei et al., 2010) was unavailable at the time of writing to do a comparison). As discussed above, two unique aspects of the proposed model are the opportunity for a sub-topic to have more than one parent, in a statistical sense, and also the lack of word duplication between scales. Figure 1 presents a zoom-in taken from Figure 5, in which the case of multiple parents is observed.

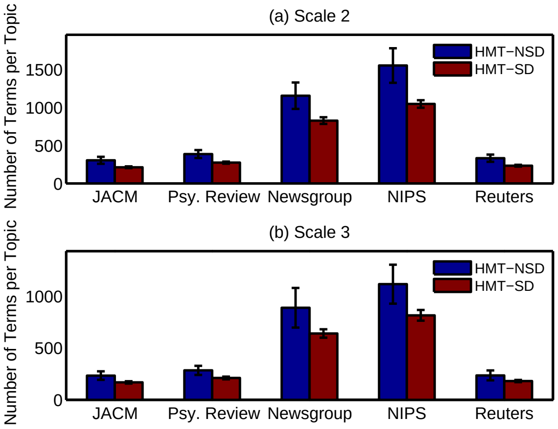

Figure 4. Number of terms per topic at scales $s = 2$ and $s = 3$ for HMT-SD and HMT-NSD.

5. Conclusions

A new hierarchical tree-based topic model has been presented. The model removes redundancies in two ways: (i) sub-topics may have multiple parents, thereby yielding a flexible, statistical branching structure; and (ii) if a word is used in a sub-topic at a particular scale, it is not re-used at deeper scales. A retrospective sampler is employed to infer both the tree depth and width (the width is scale-dependent). State-of-the-art results are achieved on five data sets, with encouraging comparisons against recently developed related models[5].

[5] This research was funded by AFOSR, ARO, NGA and ONR.

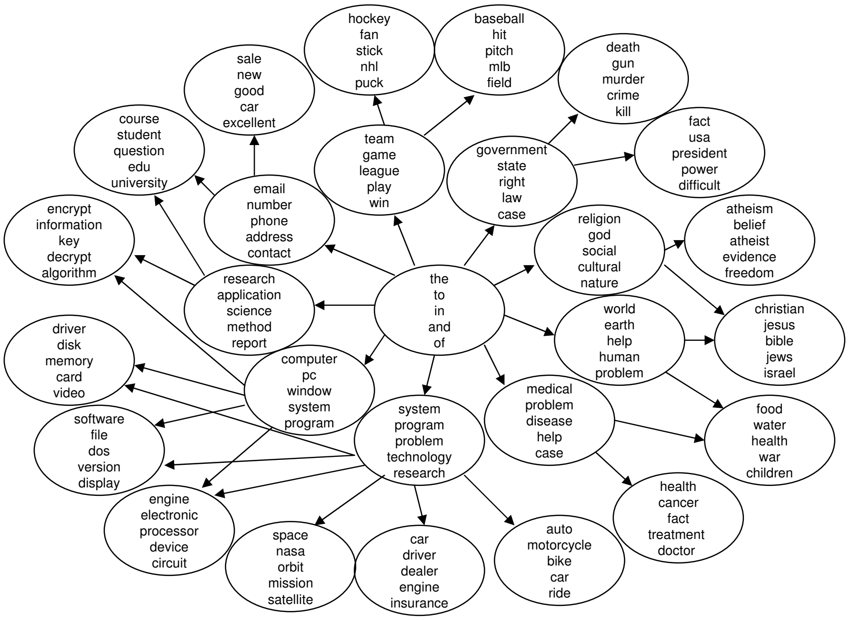
hockey fan stick nhl puck
baseball hit pitch mlb field
sale new good car excellent
death gun murder crime kill
course student question edu university
team game league play win
government state right law case
fact usa president power difficult
email number phone address contact
encrypt information key decrypt algorithm
religion god social cultural nature
atheism belief atheist evidence freedom
the to in and of
research application science method report
world earth help human problem
christian jesus bible jews israel
driver disk memory card video
computer pc window system program
system program problem technology research
medical problem disease help case
food water health war children
software file dos version display
engine electronic processor device circuit
health cancer fact treatment doctor
space nasa orbit mission satellite
car driver dealer engine insurance
auto motorcycle bike car ride

Figure 5. Example tree inferred from *20 Newsgroup* data set; these results correspond to one (typical) collection sample.

References

Adams, R. P., Ghahramani, Z., and Jordan, M. I. Tree-structured stick breaking for hierarchical data. In *Neural Information Processing Systems*, 2010.

Blei, D. M., Ng, A., and Jordan, M. I. Latent Dirichlet allocation. *Jounral of Machine Learning*, 3:993–1022, 2003.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Jouranl of the ACM*, 57(2), 2010.

Chambers, A., Smyth, P., and Steyvers, M. Learning concept graphs from text with stick-breaking priors. In *Neural Information Processing Systems*, 2010.

Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *Psychological Review*, 1(2):209–230, 1973.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. Topics in semantic representation. *Psychological Review*, 114(2):211–244, 2007.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning*, 2010.

Li, W. and McCallum, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the International Conference on Machine Learning*, 2006.

Paisley, J. and Carin, L. Hidden Markov models with stick-breaking priors. *IEEE Transactions on Signal Processing*, 57(10):3905–3917, 2009.

Papaspiliopoulos, O. and Roberts, G. O. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1): 169–186, 2008.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

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

Wang, C. and Blei, D. M. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Neural Information Processing Systems*, 2009a.

Wang, C. and Blei, D. M. Variational inference for the nested Chinese restaurant process. In *Neural Information Processing Systems*, 2009b.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. The IBP compound dirichlet process and its application to focused topic modeling. In *Proceedings of the International Conference on Machine Learning*, 2010.