Efficient Tree-Based Topic Modeling
Yuening Hu
Department of Computer Science
University of Maryland, College Park
ynhu@cs.umd.edu
Jordan Boyd-Graber
iSchool and UMIACS
University of Maryland, College Park
jbg@umiacs.umd.edu

Abstract

Topic modeling with a tree-based prior has been used for a variety of applications because it can encode correlations between words that traditional topic modeling cannot. However, its expressive power comes at the cost of more complicated inference. We extend the \textsc{SparseLDA} (Yao et al., 2009) inference scheme for latent Dirichlet allocation (LDA) to tree-based topic models. This sampling scheme computes the exact conditional distribution for Gibbs sampling much more quickly than enumerating all possible latent variable assignments. We further improve performance by iteratively refining the sampling distribution only when needed. Experiments show that the proposed techniques dramatically improve the computation time.

1 Introduction

Topic models, exemplified by latent Dirichlet allocation (LDA) (Blei et al., 2003), discover latent themes present in text collections. “Topics” discovered by topic models are multinomial probability distributions over words that evince thematic coherence. Topic models are used in computational biology, computer vision, music, and, of course, text analysis.

One of LDA’s virtues is that it is a simple model that assumes a symmetric Dirichlet prior over its word distributions. Recent work argues for structured distributions that constrain clusters (Andrzejewski et al., 2009), span languages (Jagarlamudi and Daumé III, 2010), or incorporate human feedback (Hu et al., 2011) to improve the quality and flexibility of topic modeling. These models all use different tree-based prior distributions (Section 2).

These approaches are appealing because they preserve conjugacy, making inference using Gibbs sampling (Heinrich, 2004) straightforward. While straightforward, inference isn’t cheap. Particularly for interactive settings (Hu et al., 2011), efficient inference would improve perceived latency.

\textsc{SparseLDA} (Yao et al., 2009) is an efficient Gibbs sampling algorithm for LDA based on a refactorization of the conditional topic distribution (reviewed in Section 3). However, it is not directly applicable to tree-based priors. In Section 4, we provide a factorization for tree-based models within a broadly applicable inference framework that empirically improves the efficiency of inference (Section 5).

2 Topic Modeling with Tree-Based Priors

Trees are intuitive methods for encoding human knowledge. Abney and Light (1999) used tree-structured multinomials to model selectional restrictions, which was later put into a Bayesian context for topic modeling (Boyd-Graber et al., 2007). In both cases, the tree came from WordNet (Miller, 1990), but the tree could also come from domain experts (Andrzejewski et al., 2009).

Organizing words in this way induces correlations that are mathematically impossible to represent with a symmetric Dirichlet prior. To see how correlations can occur, consider the generative process. Start with a rooted tree structure that contains internal nodes and leaf nodes. This skeleton is a prior that generates $K$ topics. Like vanilla LDA, these topics are distributions over words. Unlike vanilla LDA, their structure correlates words. Internal nodes have a distribution $\pi_{k,i}$ over children, where $\pi_{k,i}$ comes from per-node Dirichlet parameterized by $\beta_i$. Each leaf node is associated with a word, and each word must appear in at least (possibly more than) one leaf node.

To generate a word from topic $k$, start at the root. Select a child $x_0 \sim \text{Mult}(\pi_{k,\text{ROOT}})$, and traverse the tree until reaching a leaf node. Then emit the leaf’s associated word. This walk replaces the draw from a topic’s multinomial distribution over words.

\footnote{Choosing these Dirichlet priors specifies the direction (i.e., positive or negative) and strength of correlations that appear.}
The rest of the generative process for LDA remains the same, with \( \theta \), the per-document topic multinomial, and \( z \), the topic assignment.

This tree structure encodes correlations. The closer types are in the tree, the more correlated they are. Because types can appear in multiple leaf nodes, this encodes polysemy. The path that generates a token is an additional latent variable we must sample.

Gibbs sampling is straightforward because the tree-based prior maintains conjugacy (Andrzejewski et al., 2009). We integrate the per-document topic distributions \( \theta \) and the transition distributions \( \pi \). The remaining latent variables are the topic assignment \( z \) and path \( l \), which we sample jointly:

\[
p(z = k, l = \lambda | Z_-, L_-, w) \propto (\alpha_k + n_{k|d}) \prod_{(i \rightarrow j) \in \lambda} \frac{\beta_{i \rightarrow j} + n_{i \rightarrow j|k}}{\beta_{i \rightarrow j} + \sum_j \beta_{i \rightarrow j|k}}
\]

where \( n_{k|d} \) is topic \( k \)'s count in the document \( d \); \( \alpha_k \) is topic \( k \)'s prior; \( Z_- \) and \( L_- \) are topic and path distributions excluding \( w_{d,n} \); \( \beta_{i \rightarrow j} \) is the prior for edge \( i \rightarrow j \); \( n_{i \rightarrow j|k} \) is the count of edge \( i \rightarrow j \) in topic \( k \); and \( j' \) denotes other children of node \( i \).

The complexity of computing the sampling distribution is \( O(KLS) \) for models with \( K \) topics, paths at most \( L \) nodes long, and at most \( S \) paths per word type. In contrast, for vanilla LDA the analogous conditional sampling distribution requires \( O(K) \).

## 3 Efficient LDA

The SPARSELDA (Yao et al., 2009) scheme for speeding inference begins by rearranging LDA’s sampling equation into three terms:

\[
p(z = k | Z_-, w) \propto (\alpha_k + n_{k|d}) \frac{\beta + n_{w|k}}{\beta V + n_{|k}} \]

\[
\propto \frac{\alpha_k \beta}{\beta V + n_{|k}} + \frac{n_{k|d} \beta}{\beta V + n_{|k}} + \frac{(\alpha_k + n_{k|d})n_{w|k}}{\beta V + n_{|k}}
\]

Following their lead, we call these three terms “buckets”. A bucket is the total probability mass marginalizing over latent variable assignments (i.e., \( s_{LDA} \equiv \sum_k \frac{\alpha_k \beta}{\beta V + n_{|k}} \)). The three buckets are smoothing only bucket \( s_{LDA} \), document topic bucket \( r_{LDA} \), and topic word bucket \( q_{LDA} \) (we use the “LDA” subscript to contrast with our method, for which we use the same bucket names without subscripts).

Caching the buckets’ total mass speeds the computation of the sampling distribution. Bucket \( s_{LDA} \) is shared by all tokens, and bucket \( r_{LDA} \) is shared by a document’s tokens. Both have simple constant time updates. Bucket \( q_{LDA} \) has to be computed specifically for each token, but only for the (typically) few types with non-zero counts in a topic.

To sample from the conditional distribution, first sample which bucket you need and then (and only then) select a topic within that bucket. Because the topic-term bucket \( q_{LDA} \) often has the largest mass and has few non-zero terms, this speeds inference.

## 4 Efficient Inference in Tree-Based Models

In this section, we extend the sampling techniques for SPARSELDA to tree-based topic modeling. We first factor Equation 1:

\[
p(z = k, l = \lambda | Z_-, L_-, w) \propto (\alpha_k + n_{k|d})N_{k,\lambda}^{-1}(S_{\lambda} + O_{k,\lambda}).
\]

Henceforth we call \( N_{k,\lambda} \) the normalizer for path \( \lambda \) in topic \( k \), \( S_{\lambda} \) the smoothing factor for path \( \lambda \), and \( O_{k,\lambda} \) the observation for path \( \lambda \) in topic \( k \), which are

\[
N_{k,\lambda} = \prod_{(i \rightarrow j) \in \lambda} \sum_{j'} (\beta_{i \rightarrow j'} + n_{i \rightarrow j'|k})
\]

\[
S_{\lambda} = \prod_{(i \rightarrow j) \in \lambda} \beta_{i \rightarrow j}
\]

\[
O_{k,\lambda} = \prod_{(i \rightarrow j) \in \lambda} (\beta_{i \rightarrow j} + n_{i \rightarrow j|k}) - \prod_{(i \rightarrow j) \in \lambda} \beta_{i \rightarrow j}
\]

Equation 3 can be rearranged in the same way as Equation 5, yielding buckets analogous to SPARSELDA’s,

\[
p(z = k, l = \lambda | Z_-, L_-, w) \propto \frac{\alpha_k S_{\lambda}}{N_{k,\lambda}} + \frac{n_{k|d} S_{\lambda}}{N_{k,\lambda}} + \frac{(\alpha_k + n_{k|d})O_{k,\lambda}}{q_{LDA}}
\]

Buckets sum both topics and paths. The sampling process is much the same as for SPARSELDA: select which bucket and then select a topic / path combination within the bucket (for a slightly more complex example, see Algorithm 1).
Recall that one of the benefits of SPARSE LDA was that $s$ was shared across tokens. This is no longer possible, as $N_{k,\lambda}$ is distinct for each path in tree-based LDA. Moreover, $N_{k,\lambda}$ is coupled; changing $n_{i\rightarrow j|k}$ in one path changes the normalizers of all cousin paths (paths that share some node $i$).

This negates the benefit of caching $s$, but we recover some of the benefits by splitting the normalizer to two parts: the “root” normalizer from the root node (shared by all paths) and the “downstream” normalizer. We precompute which paths share downstream normalizers; all paths are partitioned into cousin sets, defined as sets for which changing the count of one member of the set changes the downstream normalizer of other paths in the set. Thus, when updating the counts for path $l$, we only recompute $N_{k,l'}$ for all $l'$ in the cousin set.

SPARSE LDA’s computation of $q$, the topic-word bucket, benefits from topics with unobserved (i.e., zero count) types. In our case, any non-zero path, a path with any non-zero edge, contributes. To quickly determine whether a path contributes, we introduce an edge-masked count (EMC) for each path. Higher order bits encode whether edges have been observed and lower order bits encode the number of times the path has been observed. For example, if a path of length three only has its first two edges observed, its EMC is $TT000000$. If the same path were observed seven times, its EMC is $TT001111$. With this formulation we can ignore any paths with a zero EMC.

Efficient sampling with refined bucket While caching the sampling equation as described in the previous section improved the efficiency, the smoothing only bucket $s$ is small, but computing the associated mass is costly because it requires us to consider all topics and paths. This is not a problem for SparseLDA because $s$ is shared across all tokens. However, we can achieve computational gains with an upper bound on $s$,

$$s = \sum_{k,\lambda} \alpha_k \prod_{(i \rightarrow j) \in \lambda} \beta_{i \rightarrow j} \prod_{(i \rightarrow j) \in \lambda} \sum_{j'} (\beta_{i \rightarrow j'} + n_{i \rightarrow j'}(k)) \leq \sum_{k,\lambda} \alpha_k \prod_{(i \rightarrow j) \in \lambda} \beta_{i \rightarrow j} \prod_{(i \rightarrow j) \in \lambda} \sum_{j'} \beta_{i \rightarrow j'} = s'. \quad (6)$$

A sampling algorithm can take advantage of this by not explicitly calculating $s$. Instead, we use $s'$ as proxy, and only compute the exact $s$ if we hit the bucket $s'$ (Algorithm 1). Removing $s'$ and always computing $s$ yields the first algorithm in Section 4.

\begin{algorithm}
\caption{SAMPLING WITH REFINED BUCKET}
1: \textbf{for} word $w$ in this document \textbf{do}
2: \hspace{1em} sample = rand() * $(s' + r + q)$
3: \hspace{1em} if sample < $s'$ \textbf{then}
4: \hspace{2em} compute $s$
5: \hspace{1em} sample = sample * $(s + r + q)/(s' + r + q)$
6: \hspace{1em} if sample < $s$ \textbf{then}
7: \hspace{2em} \textbf{return} topic $k$ and path $\lambda$ sampled from $s$
8: \hspace{1em} sample = = $s$
9: \hspace{1em} \textbf{else}
10: \hspace{2em} sample = = $s'$
11: \hspace{1em} if sample < $r$ \textbf{then}
12: \hspace{2em} \textbf{return} topic $k$ and path $\lambda$ sampled from $r$
13: \hspace{1em} sample = = $r$
14: \hspace{1em} \textbf{return} topic $k$ and path $\lambda$ sampled from $q$
\end{algorithm}

Sorting Thus far, we described techniques for efficiently computing buckets, but quickly sampling assignments within a bucket is also important. Here we propose two techniques to consider latent variable assignments in decreasing order of probability mass. By considering fewer possible assignments, we can speed sampling at the cost of the overhead of maintaining sorted data structures. We sort topics’ prominence within a document (SD) and sort the topics and paths of a word (SW).

Sorting topics’ prominence within a document (SD) can improve sampling from $r$ and $q$; when we need to sample within a bucket, we consider paths in decreasing order of $n_{k|id}$.

Sorting path prominence for a word (SW) can improve our ability to sample from $q$. The edge-masked count (EMC), as described above, serves as a proxy for the probability of a path and topic. If, when sampling a topic and path from $q$, we sample based on the decreasing EMC, which roughly correlates with path probability.

5 Experiments

In this section, we compare the running time\(^5\) of our sampling algorithm (FAST) and our algorithm with the refined bucket (RB) against the unfactored Gibbs sampler (NAIVE) and examine the effect of sorting.

Our corpus has editorials from New York Times

\(^5\)Mean of five chains on a 6-Core 2.8-GHz CPU, 16GB RAM
Average number of senses per constraint word
Average running time per iteration (S)

<table>
<thead>
<tr>
<th>Number of Topics</th>
<th>T50</th>
<th>T100</th>
<th>T200</th>
<th>T500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Naive</strong></td>
<td>5.700</td>
<td>12.655</td>
<td>29.200</td>
<td>71.223</td>
</tr>
<tr>
<td><strong>Fast</strong></td>
<td>4.935</td>
<td>9.222</td>
<td>17.559</td>
<td>40.691</td>
</tr>
<tr>
<td><strong>Fast-RB</strong></td>
<td>2.937</td>
<td>4.037</td>
<td>5.880</td>
<td>8.551</td>
</tr>
<tr>
<td><strong>Fast-RB-sD</strong></td>
<td>2.075</td>
<td>3.795</td>
<td>5.400</td>
<td>8.363</td>
</tr>
<tr>
<td><strong>Fast-RB-sW</strong></td>
<td>2.449</td>
<td>3.363</td>
<td>4.894</td>
<td>7.404</td>
</tr>
<tr>
<td><strong>Fast-RB-sDW</strong></td>
<td>2.225</td>
<td>3.241</td>
<td>4.672</td>
<td>7.424</td>
</tr>
</tbody>
</table>

Vocabulary Size

<table>
<thead>
<tr>
<th>Vocabulary Size</th>
<th>V5000</th>
<th>V10000</th>
<th>V20000</th>
<th>V30000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Naive</strong></td>
<td>4.815</td>
<td>12.351</td>
<td>28.783</td>
<td>51.088</td>
</tr>
<tr>
<td><strong>Fast</strong></td>
<td>2.897</td>
<td>9.063</td>
<td>20.460</td>
<td>38.119</td>
</tr>
<tr>
<td><strong>Fast-RB</strong></td>
<td>1.012</td>
<td>3.900</td>
<td>9.777</td>
<td>20.040</td>
</tr>
<tr>
<td><strong>Fast-RB-sD</strong></td>
<td>0.972</td>
<td>3.684</td>
<td>9.287</td>
<td>18.685</td>
</tr>
<tr>
<td><strong>Fast-RB-sW</strong></td>
<td>0.889</td>
<td>3.376</td>
<td>8.406</td>
<td>16.640</td>
</tr>
<tr>
<td><strong>Fast-RB-sDW</strong></td>
<td>0.828</td>
<td>3.113</td>
<td>7.777</td>
<td>15.397</td>
</tr>
</tbody>
</table>

Number of Correlations

<table>
<thead>
<tr>
<th>Number of Correlations</th>
<th>C50</th>
<th>C100</th>
<th>C200</th>
<th>C500</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Naive</strong></td>
<td>11.166</td>
<td>12.586</td>
<td>13.000</td>
<td>15.377</td>
</tr>
<tr>
<td><strong>Fast</strong></td>
<td>8.889</td>
<td>9.165</td>
<td>9.177</td>
<td>8.079</td>
</tr>
<tr>
<td><strong>Fast-RB</strong></td>
<td>3.995</td>
<td>4.078</td>
<td>3.858</td>
<td>3.156</td>
</tr>
<tr>
<td><strong>Fast-RB-sD</strong></td>
<td>3.660</td>
<td>3.795</td>
<td>3.593</td>
<td>3.065</td>
</tr>
<tr>
<td><strong>Fast-RB-sW</strong></td>
<td>3.272</td>
<td>3.363</td>
<td>3.308</td>
<td>2.787</td>
</tr>
<tr>
<td><strong>Fast-RB-sDW</strong></td>
<td>3.026</td>
<td>3.241</td>
<td>3.091</td>
<td>2.627</td>
</tr>
</tbody>
</table>

Table 1: The average running time per iteration (S) over 100 iterations, averaged over 5 seeds. Experiments begin with 100 topics, 100 correlations, vocab size 10000 and then vary one dimension: number of topics (top), vocabulary size (middle), and number of correlations (bottom).

from 1987 to 1996. Since we are interested in varying vocabulary size, we rank types by average tf-idf and choose the top \( V \). WordNet 3.0 generates the correlations between types. For each synset in WordNet, we generate a subtree with all types in the synset—that are also in our vocabulary—as leaves connected to a common parent. This subtree’s common parent is then attached to the root node.

We compared the **Fast** and **Fast-RB** against **Naive** (Table 1) on different numbers of topics, various vocabulary sizes and different numbers of correlations. **Fast** is consistently faster than **Naive** and **Fast-RB** is consistently faster than **Fast**. Their benefits are clearer as distributions become sparse (e.g., the first iteration for **Fast** is slower than later iterations). Gains accumulate as the topic number increases, but decrease a little with the vocabulary size. While both sorting strategies reduce time, sorting topics and paths for a word (sW) helps more than sorting topics in a document (sD), and combining the

two is (with one exception) better than either alone.

As more correlations are added, **Naive**’s time increases while that of **Fast-RB** decreases. This is because the number of non-zero paths for uncorrelated words decreases as more correlations are added to the model. Since our techniques save computation for every zero path, the overall computation decreases as correlations push uncorrelated words to a limited number of topics (Figure 1). Qualitatively, when the synset with “king” and “baron” is added to a model, it is associated with “drug, inmate, colombia, waterfront, baron” in a topic; when “king” is correlated with “queen”, the associated topic has “king, parade, museum, queen, jackson” as its most probable words. These represent reasonable disambiguations. In contrast to previous approaches, inference speeds up as topics become more semantically coherent (Boyd-Graber et al., 2007).

6 Conclusion

We demonstrated efficient inference techniques for topic models with tree-based priors. These methods scale well, allowing for faster exploration of models that use semantics to encode correlations without sacrificing accuracy. Improved scalability for such algorithms, especially in distributed environments (Smola and NarayanaMurthy, 2010), could improve applications such as cross-language information retrieval, unsupervised word sense disambiguation, and knowledge discovery via interactive topic modeling.

---

\(^6\)13284 documents, 41554 types, and 2714634 tokens.
Acknowledgments

We would like to thank David Mimno and the anonymous reviewers for their helpful comments. This work was supported by the Army Research Laboratory through ARL Cooperative Agreement W911NF-09-2-0072 and NSF grant #1018625. Any opinions or conclusions expressed are the authors’ and do not necessarily reflect those of the sponsors.

References


