Dynamic Topic Model with Hierarchical Two-parameter Poisson-Dirichlet Process

Jinjing Li$^1$  Wray Buntine$^2$

$^1$ANU (u5213689)

$^2$NICTA (Supervisor)

Project Presentation, November 2013
Outline

1. Introduction
2. Proposed New Model
3. Challenges and Innovations
4. Experiments
5. Conclusion
Increasing amount of text data available, difficult to find and discover what we are looking for

Topic model algorithms can be used for discovering the themes of large unstructured collection of documents
  - Techniques for finding patterns of words using non-parametric probabilistic models

Starting point: Latent Dirichlet Allocation Model (LDA, Blei et al. 2003)
A Review of LDA

Figure: Graphical representation of the LDA Model

Generative Process of Latent Dirichlet Allocation (LDA) model

1. For each document
2. Draw $\tilde{\theta}_d \sim \text{Dir} \left(a, \tilde{\theta}\right)$,
   
   $\tilde{\phi}_k \sim \text{Dir} \left(b, \tilde{\phi}\right)$
3. For each word:
   1. Draw $z_{dl} \sim \text{Mult} \left(\tilde{\theta}_d\right)$
   2. Draw $w_{dl} \sim \text{Mult} \left(\tilde{\phi}_z\right)$
Drawbacks of LDA

- Static assumption
  - May not be appropriate for documents with time attributes, e.g. journals, emails, news etc.

- Alternative modelling method,
  - DTM (Blei and Lafferty, 2006): Chain the parameters using Gaussian process (mostly pragmatic reason)
  - TOT (McCallum, 2006): assuming a static topic-word distribution
Proposed New Dynamic Topic Model

- Adapts a variant of Hierarchical Dirichlet Process (HDP) based model as it is regarded as a better version of original LDA
- Allows dynamic evolution for both topic and topic-word distributions
- Differentiates the evolution of different topics
- Potential capacity to scale
Graphical Representation of the Proposed Model

Data Generation Process

1. Draw topic mixtures
   \( \tilde{\mu}_t \sim \text{PDP} \left( a_\mu, b_\mu, \tilde{\mu}_{t-1} \right) \)

2. Draw topics
   \( \tilde{\phi}_{tk} \sim \text{PDP} \left( a_\phi, b_{\phi k}, \tilde{\phi}_{t-1,k} \right) \)

3. For each document
   1. Draw \( \tilde{\theta}_{td} \sim \text{PDP} \left( a_\theta, b_\theta, \tilde{\mu}_t \right) \)
   2. For each word:
      1. Draw \( z_{tdl} \sim \text{Mult} \left( \tilde{\theta}_{td} \right) \)
      2. Draw \( w_{tdl} \sim \text{Mult} \left( \tilde{\phi}_{tz} \right) \)

PDP: Poisson-Dirichlet Process (A generalised Dirichlet process, a.k.a Pitman-Yor Process)
Challenges of the New Model

- Intractable posterior distribution in hierarchical setting
- Hyper-parameter estimations
- Practical computation issues
Challenges of the New Model

- Intractable posterior distribution in hierarchical setting
- Hyper-parameter estimations
- Practical computation issues
Inference

- Intractable posterior distribution raises the need of approximate inference techniques

- Use of “Auxiliary variable” trick / “Table indicator” sampling technique (Robert and Casella, 2004; Chen et al., 2011; Buntine and Hutter, 2012) for hierarchical PDP sampling

- If $\bar{\theta}$ is the parent distribution, and $z$ (topic) follows $\text{Discrete}_K(\bar{\theta})$, the marginalised likelihood of $\bar{z}$ with auxiliary variable

$$p(\bar{z}, \bar{c} | a, b, \bar{\theta}, \text{PDP}) = \frac{(b|a)c}{(b)_N} \prod_k S_{c_k,a}^{n_k} (\theta_k)^{c_k}$$

where $a$ and $b$ are PDP parameters, $c_k$ and $n_k$ are the “table count” and the “number of customers” (Chinese Restaurant Process notation), $(\cdot)_N$ refers to Pochhammer symbol. $S$ notation refers to generalised Stirling number.
Inference and Sampling

- Two sets of auxiliary variables and indicators as we allow dynamic evolution from both topic-document side and the topic-word side of the model.

- The marginalised likelihood is a function of indicators \((r, r')\), which records how far back the word travels back in time (epoch).

- Sample \(p(z, r, r')\) using a block Gibbs sampler (please refer to the paper for equations).

- Get predicted topic \(p(z)\) by marginalising out \(r\) and \(r'\).
Challenges of the New Model

- Intractable posterior distribution in hierarchical setting
- Hyper-parameter estimations
- Practical computation issues
Hyper-parameters

- The new models uses a number of distributions and allows heterogeneous evolution of topic-word distributions
  - More hyper-parameters means more assumptions are required
- Adaptive Rejection Sampling (Gilks and Wild, 1992) of hyper-parameters for every specified cycle of Gibbs sampling
Challenges of the New Model

- Intractable posterior distribution in hierarchical setting
- Hyper-parameter estimations
- Practical computation issues
Computation Efficiencies

- The model is complicated with many operations that requires computations over time (epoch)
- Text corpora have some common properties, e.g. bounded number of words, frequent numeric repetitions etc.
- Extensive use of caches in the computation
  - Rewrite the sampling equations into recursive functions
  - Caching the intermediate results which will likely be used later
  - Cache the Stirling number calculation
  - Best case scenario (most common) $O(t^2) \rightarrow O(1)$
- C Implementation
Experiments

- Dataset (subset of ABC News Corpus)
  - News with word “obesity” (OB) in the past decade (63k words)
  - News with word “Japan” (JP) in the past decade (617k words)
- Pre-processing
  - Rare words and stop words removed
Validate the model structure, basic computation and compare with LDA

<table>
<thead>
<tr>
<th>Corpus</th>
<th>OB1-LDA</th>
<th>OB1-NEW</th>
<th>JP1-LDA</th>
<th>JP1-NEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iteration 50</td>
<td>10.96</td>
<td>10.91</td>
<td>12.40</td>
<td>12.15</td>
</tr>
<tr>
<td>Iteration 100</td>
<td>10.86</td>
<td>10.75</td>
<td>12.27</td>
<td>12.03</td>
</tr>
<tr>
<td>Iteration 150</td>
<td>10.79</td>
<td>10.70</td>
<td>12.23</td>
<td>12.01</td>
</tr>
<tr>
<td>Iteration 200</td>
<td>10.74</td>
<td>10.69</td>
<td>12.20</td>
<td>12.00</td>
</tr>
</tbody>
</table>

Table: Perplexity comparison for single epoch corpora
Experiment Results (Multi Epochs)

Validate computation in dynamic settings and hyper-parameter sampling

<table>
<thead>
<tr>
<th>Corpus</th>
<th>JP10</th>
<th>JP40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyper-parameters</td>
<td>Fixed</td>
<td>Sampled</td>
</tr>
<tr>
<td>Iteration 50</td>
<td>13.38</td>
<td>12.96</td>
</tr>
<tr>
<td>Iteration 100</td>
<td>13.16</td>
<td>12.76</td>
</tr>
<tr>
<td>Iteration 150</td>
<td>13.08</td>
<td>12.69</td>
</tr>
<tr>
<td>Iteration 200</td>
<td>13.03</td>
<td>12.63</td>
</tr>
</tbody>
</table>

Table: Perplexity comparison for multi-epoch corpora
Experiment Result (Hyper-Parameter Sampling)

A distribution of sampled hyper-parameters for word-topic distribution $PDP(a_\phi, b_{\phi k})$

Figure: Distribution of $b_\phi$ in OB10 corpus
The impact of $b\phi$ on the originated epoch of words and the effectiveness of the caching system

**Figure**: Words originated from previous epochs
The project has

- Developed a new dynamic topic model with some theoretical advantages
- Derived and implemented the Gibbs sampling algorithm used in the model
- Developed a caching system which improves the sampling speed
- Demonstrated the performance of the new model and the cache design
- Implemented the new algorithm in C (under Mozilla Public License)