Experiments with Non-parametric Topic Models

Swapnil Mishra $^1$ and Wray Buntine $^2$

$^1$NICTA + ANU

$^2$Monash University

Tuesday 26th August, 2014
We are hiring!

Melbourne, world’s most livable city, again

junior faculty in data science
Outline

1. Background
2. The Non Parametric Topic Model
3. Experiments
4. Conclusion
Topic Models

- *Topic Models* discover hidden themes in text data to aid understanding
  - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003)
- recent research develops higher performance topic models
Topic Models

- **Topic Models** discover hidden themes in text data to aid understanding
  - Latent Dirichlet Allocation Model (LDA, Blei et al. 2003)
- recent research develops higher performance topic models
- but why should you care?
Using Topic Models

topic models are the leading edge of a new wave of deep latent semantic models applied to real NLP tasks:

e.g., document segmentation, word sense disambiguation, facet discovery, ...

in the middle of this segmentation model is a topic model

- better topic models are important components
High Fidelity Topic Models and Visualisation

Can get 100’s of topics from 1000’s of documents providing real insight.

interface from MetaHeuristica.com
Background

Text and Burstiness

Original news article:

Women may only account for 11% of all Lok-Sabha MPs but they fared better when it came to representation in the Cabinet. Six women were sworn in as senior ministers on Monday, accounting for 25% of the Cabinet. They include Swaraj, Gandhi, Najma, Badal, Uma and Smriti.

Bag of words:

11% 25% Badal Cabinet(2) Gandhi Lok-Sabha MPs Monday Najma Six Smriti Swaraj They Uma Women account accounting all and as better but came fared for(2) in(2) include it may ministers of on only representation senior sworn the(2) they to were when women

NB. “Cabinet” appears twice! It is **bursty**

(see Doyle and Elkan 2009)
Previous Work

EXTENDING


“Topic models with power-law using Pitman-Yor process,” Sato and Nakagawa 2010


COMPARING AGAINST
Better Sampling Methods for HDP and HPYP

Sampling for hierarchical Dirichlet Processes and Pitman-Yor Processes:

The Old: hierarchical Chinese Restaurant Processes (CRP) from Teh et al. 2006.


- requires no dynamic memory
- more rapid mixing so leads to better models
- more easily applied to more complex models
- demonstrated extensively on different problems!
Outline

1 Background

2 The Non Parametric Topic Model
   - Evolution of Models
   - Our Non-parametric Topic Model

3 Experiments

4 Conclusion
Evolution of Models

LDA - Scalar
original LDA
The Non Parametric Topic Model

Evolution of Models

\[ \vec{\alpha} \]
\[ \vec{\theta}_d \]
\[ (z_{d,n}) \]
\[ (w_{d,n}) \]
\[ \vec{\beta} \]
\[ \vec{\phi}_k \]
\[ N \]
\[ D \]
\[ K \]

LDA - Vector
adds asymmetric Dirichlet prior like Wallach et al.;
is also truncated HDP-LDA;
implemented by Mallet as assymetric-symmetric LDA
HDP adds proper modelling of topic prior like Teh et al.
The Non Parametric Topic Model

Evolution of Models

NP-LDA
adds power law on word distributions like Sato et al. and estimation of background word distribution
Evolution of Models

The Non Parametric Topic Model

NP-LDA with Burstiness
add’s burstiness like Doyle and Elkan
Our Non-parametric Topic Model

\[ \{\tilde{\theta}_d\} = \text{document} \otimes \text{topic matrix} \]

\[ \{\tilde{\phi}_k\} = \text{topic} \otimes \text{word matrix} \]

- Full fitting of priors on topic \( \otimes \) word and document \( \otimes \) topic matrices (red nodes).
- Topic \( \otimes \) word vectors \( \tilde{\phi}_k \) specialised to the document to yield \( \tilde{\psi}_k \).
- This models burstiness (blue node).
Handling Burstiness

- proposed by Doyle and Elkan 2009
- used a slow variational method
- we developed a Gibbs sampler that acts as a front end to any LDA-style model with Gibbs, e.g.
  - dynamic topic models
  - bibliographic network models
- implemented as a C function that calls the Gibbs sampler
- adds smallish memory (20%) and time (30%) overhead
Outline

1. Background
2. The Non Parametric Topic Model
3. Experiments
   - Runtime
   - Performance
4. Conclusion
## Runtime Characteristics

<table>
<thead>
<tr>
<th>Alg.</th>
<th>mins.</th>
<th>Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>11</td>
<td>630</td>
</tr>
<tr>
<td>Burst LDA</td>
<td>20</td>
<td>690</td>
</tr>
<tr>
<td>HDP-LDA</td>
<td>20</td>
<td>760</td>
</tr>
<tr>
<td>Burst HDP-LDA</td>
<td>30</td>
<td>850</td>
</tr>
<tr>
<td>NP-LDA</td>
<td>35</td>
<td>840</td>
</tr>
<tr>
<td>Burst NP-LDA</td>
<td>45</td>
<td>930</td>
</tr>
<tr>
<td>Online HDP</td>
<td>236</td>
<td>1800</td>
</tr>
</tbody>
</table>

Cycle times and memory requirements on the LA Times TREC 4 data.

- “Burst” is the burstiness version,
- “NP-LDA” is our default non-parametric version with full sampling of hyperparameters.
- “Online HDP” is (Wang, Paisley and Blei) in Python. Recent C++ version from Wang “faster”.
Performance Metrics

Perplexity:  
- measure of test set likelihood;
- equal to effective size of vocabulary;
- we use “document completion,”
- see Wallach, Murray, Salakhutdinov, and Mimno, 2009.

PMI:  
- measure of topic coherence: “average pointwise mutual information between all pairs of the top 10 words in the topic”
Performance on Reuters-21578 ModLewis Split

Training on 11314 news articles with vocabulary of 16994.
Comparison with Mallet

Asymmetric-symmetric option by Mimno implements Wallach’s method for estimating $\tilde{\alpha}$ since 2008. Great truncated HDP-LDA implementation.

**NB.** previous HDP-LDA work does not compare with AS-LDA in Mallet because they didn’t realise it was effectively truncated HDP-LDA.
Comparison with Sato, Kurihawa, Nakagawa, KDD 2012

By our analysis, the best performing variational algorithm (without split-merge).

**NB.** WSJ3000 has 3000 docs and vocab with 30000 words! Thanks to Issei Sato for providing the data.
Comparison to Bryant + Sudderth (2012) on NIPS data

Experiments

Performance

Buntine + Mishra (Monash + ANU)

NP-TMs

Tuesday 26th August, 2014 20 / 24
Effect of Hyperparameters on the Number of Topics

![Graph showing the effect of hyperparameters on the number of topics.](image)

- Test Perplexity vs. Topics
- PMI of Topics vs. Topics

Hyperparameters:
- beta=0.001
- beta=0.01
- beta=0.1
- beta=0.5
- beta=sampled

Buntine + Mishra (Monash + ANU)
NP-TMs
Tuesday 26th August, 2014
Outline

1 Background

2 The Non Parametric Topic Model

3 Experiments

4 Conclusion
Conclusion

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this

burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
Conclusion

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)
Conclusion

- NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
- burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
- NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)
- hyper-parameter adaptation is important for a high performance model
NP-LDA model is about 50% slower than HDP-LDA but usually performs substantially better
  - most previous work failed to show this
burstiness has substantially improved topic comprehensibility and dramatically improved perplexity
  - now available, efficiently, for a broad variety of models
NP-LDA using block table indicator Gibbs sampling methods from Chen et al. (2011) are superior to (several) state-of-the-art variational algorithms (when not online)
hyper-parameter adaptation is important for a high performance model
speedup efficiency of 70%-75% has been achieved running with 6 parallel threads on multi-core machines using atomic operations
Thank You

Questions?

- Grab our code from
  
  https://github.com/wbuntine/topic-models or
  
  http://mloss.org/software/view/527/

- thanks to Issei Sato and David Mimno for data and discussion

- special thanks to next paper (Aaron Li et al.) for showing us how to speed up our algorithm again!