Applying hLDA to Practical Topic Modeling

Joseph Heng
lengerfulluse@gmail.com

CIST Lab of BUPT

March 17, 2013

Outline

1. Introduction
   - HLDA
   - Discussion
2. Bayesian Clue
   - the nested CRP
   - GEM Distribution
   - Dirichlet Distribution
   - Posterior Inference

Outline

Outline

Outline

Background

♣ **The Goal in Our Paper.**
HLDA has been proved to be a powerful tool. One of the bottlenecks which prevent its large-scale application is that we cannot find a quick and effective approach to modeling new data properly. There exist lots of factors, eg. hyper-parameter settings, uncertainty of random algorithms and features of different corpus.

♣ **Probabilistic Topic Models**
Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.

♣ **LDA-related Introduction**
LDA and other topic models are part of the larger field of probabilistic modeling. In generative probabilistic modeling, we treat our data as arising from a generative process that includes hidden variables.

Merits of HLDA

1. Generative process for documents [2]
2. Posterior approximate inference with Gibbs sampling [1]

$$p(z_{d,n}|z_{-(d,n)}, c, w, \pi, \eta) \propto p(z_{d,n}|z_{-d,-n}, m, \pi) p(z_{d,n}|z, c, w_{-(d,n)}, \eta) \quad (1)$$

$$p(c_d|w, c_{-d}, z, \eta, \gamma) \propto p(c_d|c_{-d}, \gamma) p(w_d|c, w_{-d}, z, \eta) \quad (2)$$

Merits of HLDA

1. Generative process for documents [2]
2. Posterior approximate inference with Gibbs sampling [1]

$$p(z_{d,n}|z_{-(d,n)}, c, w, \pi, \eta) \propto \\ p(z_{d,n}|z_{-d,-n}, m, \pi)p(z_{d,n}|z, c, w_{-(d,n)}, \eta) \quad (1)$$

$$p(c_d|w, c_{-d}, z, \eta, \gamma) \propto p(c_d|c_{-d}, \gamma)p(w_d|c, w_{-d}, z, \eta) \quad (2)$$

3. Assessing convergence and approximating the mode.

$$L^{(t)} = \log p(c_{1:D}^{(t)}, z_{1:D}^{(t)}, z_{1:D}|\eta, \gamma, m, \pi) \quad (3)$$

Merits of HLDA

1. Generative process for documents [2]
2. Posterior approximate inference with Gibbs sampling [1]

$$p(z_{d,n}|z_{-(d,n)}, c, w, \pi, \eta) \propto p(z_{d,n}|z_{-d,-n}, m, \pi)p(z_{d,n}|z, c, w_{-(d,n)}, \eta) \quad (1)$$

$$p(c_d|w, c_{-d}, z, \eta, \gamma) \propto p(c_d|c_{-d}, \gamma)p(w_d|c, w_{-d}, z, \eta) \quad (2)$$

3. Assessing convergence and approximating the mode.

$$L^{(t)} = \log p(c_{1:D}^{(t)}, z_{1:D}^{(t)}, z_{1:D}|\eta, \gamma, m, \pi) \quad (3)$$

Practical Difficulty

- ♣ **What's practical problem [3] when using hDLA to topic modeling?**
- ♣ **Why unified framework?**

Practical Difficulty

- ♣ **What's practical problem [3] when using hDLA to topic modeling?**
- ♣ **Why unified framework?**

Practical Difficulty

♣ **What's practical problem [3] when using hDLA to topic modeling?**

♣ **Why unified framework?**

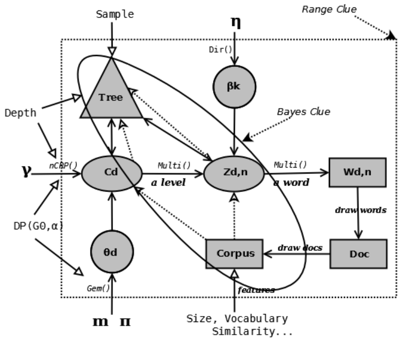


Figure : unified analysis framework with two clues

Generative Process

Document are assumed to be drawn from the following process.

- For each node k $\in$ T in the infinite tree, draw a topic $\beta_k \sim Dirichlet(\gamma)$.
- For each document, d $\in$ 1,2,...,D
  - Draw $C_d \in nCRP(\gamma)$ to choose the path
  - Draw a distribution over levels in the tree, $\theta_d|m, \pi \in GEM(m, \pi)$.
  - For each word,
    - Choose level $Z_{d,n}|\theta \in Mult(\theta_d)$.
    - Choose word $W_{d,n}|Z_{d,n}, C_d, \beta \in Mult(\beta_{c_d}, [Z_{d,n})$.

nCRP

- Introduction to CRP algorithm
- γ parameters
- Experiments with γ
- Comparision

GEM

- The Different View of DP
- Parameters m and π
- Experiments with m and π

Dirichlet Process

1. Experiment with parameter
2. Relationship with Three above.

Iterator Convergency

- ♣ Monte Carlo Markov Chain
- ♣ (Collapsed) Gibbs Sampling Algorithm

Global Modeling Strategy

- **Tree Depth**

Global Modeling Strategy

- **Tree Depth**

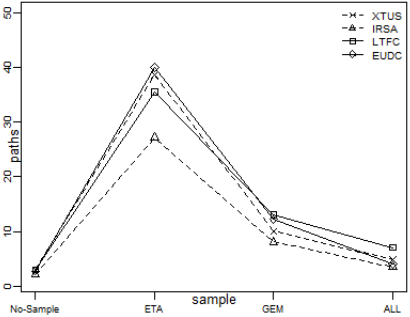
theme\depth	3	4	5
XTUS	4.9	59.7	78.7
IRSA	3.6	35.8	46.5
LTFC	7.1	54.2	74.8
EUDC	4.1	59.2	83.5
GBAB	3.1	47	66.2

Global Modeling Strategy

- **Tree Depth**
- **Sampling or not**

Global Modeling Strategy

- **Tree Depth**
- **Sampling or not**

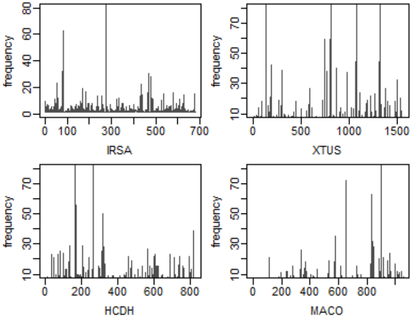


Global Modeling Strategy

- **Tree Depth**
- **Sampling or not**
- **Corpus features**

Global Modeling Strategy

- **Tree Depth**
- **Sampling or not**
- **Corpus features**



Manual Modeling Procedure

- *Generate hLDA input and extract features from corpus.*
- *Approximate depth of tree.*
- *Topic parameter for each level.*
- *nCRP parameter for non-leaf levels.*
- *m*, π *parameter for words allocations.*
- *Sampling for hyper-parameters or not.*

Empirical Results

Experiments have been conducted with the guide of modeling procedure above with only three modifications to the settings.

theme	level#1	level#2	hLDA#1	hLDA#2	score
XTUS	4	9	5.3	10	4
EUDC	4	5	5.2	9	3
IRSA	2	5	3.9	8	4
HCDH	5	7	4.8	10	4
CQWF	4	9	5.3	10	5
SBAG	4	10	6	10	5
GBAB	6	8	4.8	11	3
LTFC	6	14	7.2	12	5
MACO	4	7	6.7	8	4
NOHN	5	9	7	9	4

Reference

Asli C and Dilek H.
A hybrid hierarchical model for multi-document summarization.
ACL 10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 815–824, 2010.

Blei David, Andrew Ng, and Jordan.
Latent dirichlet allocation.
Journal of Machine Learning Research, 2003.

Paisley J, Wang C, Blei D M, and Jordan.
Nested hierarchical dirichlet proceses.
arXiv preprint arXiv:1210.6738, 2012.