Bibliographic Analysis with the Citation Network Topic Model

MONASH University | ANU THE AUSTRALIAN NATIONAL UNIVERSITY | NICTA

Kar Wai Lim (ANU & NICTA), Wray Buntine (Monash University)

HIGHLIGHT

Background: Applying topic models to bibliographic data is challenging due to the non-trivial combination of text and network modelling. Jointly modelling text and network with a full Bayesian model also leads to complicated learning algorithms which are not easy to implement.

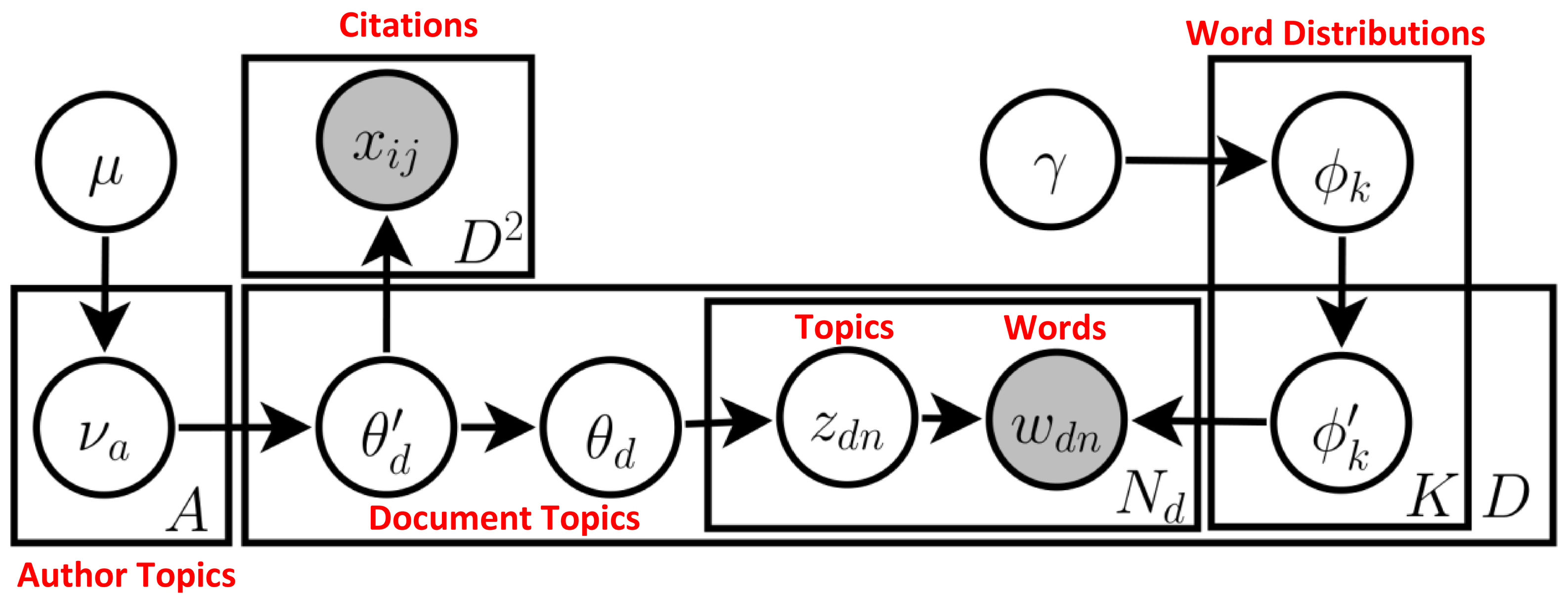
Objectives: Propose a topic model for bibliographic data that

1. Jointly models text and network, as well as additional metadata (*e.g.* authors).
2. Gives a learning algorithm that is simple to implement and understand.
3. Provides comprehensive qualitative results for corpus exploration.

Contributions: Citation Network Topic Model

1. A full Bayesian non-parametric topic model for bibliographic data.
2. Models text with a hierarchical Pitman-Yor process (HPYP) topic model.
3. Models citations with a Poisson network model.
4. Gives a novel learning algorithm that marginalises out probability distributions, treating text and citation data as counts.
5. Outperforms previous work on model fitting and clustering.
6. Facilitates various qualitative analysis such as authors and topics analysis.

CITATION NETWORK TOPIC MODEL (CNTM)



Characteristics:

1. Authors' topics influence documents' topics.
2. Document topics hierarchy reflects the difference in text and citations.
3. Word burstiness is modelled following Buntine and Mishra (2014).

Full model:

$$\begin{aligned}
\mu &\sim \mathrm{GEM}(\alpha^{\mu}, \beta^{\mu}) \\
\nu_a | \mu &\sim \mathrm{PYP}(\alpha^{\nu_a}, \beta^{\nu_a}, \mu) \\
\theta'_d | a_d, \nu &\sim \mathrm{PYP}(\alpha^{\theta'_d}, \beta^{\theta'_d}, \nu_{a_d}) \\
\theta_d | \theta'_d &\sim \mathrm{PYP}(\alpha^{\theta_d}, \beta^{\theta_d}, \theta'_d) \\
\gamma &\sim \mathrm{PYP}(\alpha^{\gamma}, \beta^{\gamma}, H^{\gamma}) \\
\phi_k | \gamma &\sim \mathrm{PYP}(\alpha^{\phi_k}, \beta^{\phi_k}, \gamma)
\end{aligned}$$

$$\begin{aligned}
\phi'_{dk} | \phi_k &\sim \mathrm{PYP}(\alpha^{\phi'_{dk}}, \beta^{\phi'_{dk}}, \phi_k) \\
z_{dn} | \theta_d &\sim \mathrm{Discrete}(\theta_d) \\
w_{dn} | z_{dn}, \phi'_d &\sim \mathrm{Discrete}(\phi'_{d z_{dn}}) \\
x_{ij} | \lambda_{ij} &\sim \mathrm{Poisson}(\lambda_{ij}) \\
\lambda_{ij} &= \lambda_i^+ \lambda_j^- \sum_k \lambda_k^T \theta'_{ik} \theta'_{jk}
\end{aligned}$$

MODEL REPRESENTATION

Count representation: Each probability vector (ν, θ *etc.*) is marginalised out and information is stored as counts. The probability vector, in Chinese Restaurant Process terminology, is represented by integers known as customer counts c and table counts t. Its explicit probability vector can be recovered from these counts.

Modularised representation: CNTM's posterior likelihood can be broken down into product of marginalised posterior of the individual probability vectors. The marginalised posterior of each probability vector only depends on their counts and hyperparameters.

Modularised posterior:

$$f(\mathcal{N}) = \frac{(\beta^{\mathcal{N}} | \alpha^{\mathcal{N}})_{T^{\mathcal{N}}}}{(\beta^{\mathcal{N}})_{C^{\mathcal{N}}}} \prod_k S^{c_k^{\mathcal{N}}}_{t_k^{\mathcal{N}}, \alpha^{\mathcal{N}}} \qquad \text{for } \mathcal{N} \sim \mathrm{PYP}(\alpha^{\mathcal{N}}, \beta^{\mathcal{N}}, \mathcal{P})$$

Posterior for HPYP topic model:

$$f(\mu) \left(\prod_{a=1}^{A} f(\nu_a) \right) \left(\prod_{d=1}^{D} f(\theta'_d) \, f(\theta_d) \prod_{k=1}^{K} f(\phi'_{dk}) \right) \left(\prod_{k=1}^{K} f(\phi_k) \right) f(\gamma) \left(\prod_v \left(\frac{1}{|\mathcal{V}|} \right)^{t_v^{\gamma}} \right)$$

Posterior for Poisson network model:

$$p(\mathbf{X} | \lambda, \theta') = \left(\prod_i (\lambda_i^+)^{g_i^+} (\lambda_i^-)^{g_i^-} \right) \prod_{ij} \left(\sum_k \lambda_k^T \theta'_{ik} \theta'_{jk} \right)^{x_{ij}} \exp\left(- \sum_{ijk} \lambda_i^+ \lambda_j^- \lambda_k^T \theta'_{ik} \theta'_{jk} \right)$$

INFERENCE TECHNIQUES

Collapsed Gibbs sampler for HPYP topic model:

1. The algorithm is analogous to LDA's Gibbs sampler.
2. First decrements a word and its associated customer counts c and table counts t.
3. Then jointly samples both topic assignments z and the associated counts.

Metropolis-Hastings algorithm for Poisson network:

1. Introduces an auxiliary variable to denote the citing topic that causes a citation.
2. Assumes there is only one citing topic for each citation (reasonable in practice).
3. This simplifies the posterior of the Poisson network.

$$p(\mathbf{X}, \mathbf{Y} | \lambda, \theta') \propto \prod_i (\lambda_i^+)^{g_i^+} (\lambda_i^-)^{g_i^-} \prod_k \left(\lambda_k^T \right)^{\frac{1}{2} \sum_i h_{ik}} \prod_{ik} {\theta'_{ik}}^{h_{ik}} \exp\left(- \sum_{ij} \lambda_i^+ \lambda_j^- \lambda_{y_{ij}}^T \theta'_{i y_{ij}} \theta'_{j y_{ij}} \right)$$

4. Note the probability vector θ' in the posterior can be marginalised out.
5. This contributes additional counts to the existing customer and table counts.
6. The MH algorithm is straight forward and is similar to the Gibbs sampler.
7. High acceptance probabilities.

Hyperparameter sampling:

1. Auxiliary variable sampler for PYP hyperparameters.
2. Gibbs sampler for the Poisson network hyperparameters.

DATA

Research publications: 3 corpus queried from CiteSeer^X and 3 existing corpus.

Datasets	Publications	Citations	Authors	Vocabulary	Words/Doc	%Repeat
1. ML	139 227	1 105 462	43 643	8 322	59.4	23.3
2. M10	10 310	77 222	6 423	2 956	57.8	24.3
3. AvS	18 720	54 601	11 898	4 770	58.9	17.0
4. AI	3 312	4 608	–	3 703	31.8	–
5. Cora	2 708	5 429	–	1 433	18.2	–
6. PubMed	19 717	44 335	–	4 209	67.6	40.1

RESULTS

Quantitative evaluations: See paper for goodness of fit and clustering results.

Qualitative analysis:

1. Topic exploration from a collection of publication data. (ML dataset)

Topic	Top Words
Reinforcement Learning	reinforcement, agents, control, state, task
Object Recognition	face, video, object, motion, tracking
Data Mining	mining, data mining, research, patterns, knowledge
SVM	kernel, support vector, training, clustering, space
Speech Recognition	recognition, speech, speech recognition, audio, hidden markov

2. Analyse authors' interest from the author-topic distribution. (M10 dataset)

Author	Topic	Top Words
D. Aerts	Quantum Theory	quantum, theory, quantum mechanics, classical, quantum field
Y. Bengio	Neural Network	networks, learning, recurrent neural, neural networks, models
C. Boutilier	Decision Making	decision making, agents, decision, theory, agent
S. Thrun	Robot Learning	robot, robots, control, autonomous, learning
M. Baker	Financial Market	market, risk, firms, returns, financial

3. Visualise author-topics relationships. (ML dataset)