Bibliographic Analysis with the Citation Network Topic Model

Kar Wai Lim[1], Wray Buntine[2]

[1]The Australian National University & NICTA
Canberra, Australia

[2]Monash University
Melbourne, Australia

MONASH University

ANU
THE AUSTRALIAN NATIONAL UNIVERSITY

NICTA

Outline

Introduction

- Bibliographic data:
  - Journal publications, conference papers, online articles *etc.*
  - Accompanied by metadata: authors, keywords, citations...
  - Grow exponentially – the need to analyse them automatically!

Introduction

- Bibliographic data:
  - Journal publications, conference papers, online articles *etc.*
  - Accompanied by metadata: authors, keywords, citations...
  - Grow exponentially – the need to analyse them automatically!
- Topic models:
  - Unsupervised Bayesian methods for thematic exploration in documents.
  - Allow automatic topics extraction from large text corpus.
  - Example:
    - HDP-LDA (Teh and Jordan, 2010) – Bayesian model with nonparametric priors.
    - Author-topic model (Rosen-Zvi et al., 2004) – incorporate author info.

Introduction

- Bibliographic data:
  - Journal publications, conference papers, online articles *etc.*
  - Accompanied by metadata: authors, keywords, citations...
  - Grow exponentially – the need to analyse them automatically!
- Topic models:
  - Unsupervised Bayesian methods for thematic exploration in documents.
  - Allow automatic topics extraction from large text corpus.
  - Example:
    - HDP-LDA (Teh and Jordan, 2010) – Bayesian model with nonparametric priors.
    - Author-topic model (Rosen-Zvi et al., 2004) – incorporate author info.
- Network models:
  - Model links (connections) between two items given some features.
  - Special case: Citation models for documents.
  - Example:
    - Mixed topic-link models (Zhu, Yan, Getoor, Moore, 2013) – model citation link with topic tag.

Motivation

- Make use of available metadata for bibliographic analysis on publication data.
- Explore thematic structure of publication corpus and study authors' influence.
- Later:
  - estimate individual author contribution,
  - tag citation with explanation.

Networks of Probability Vectors

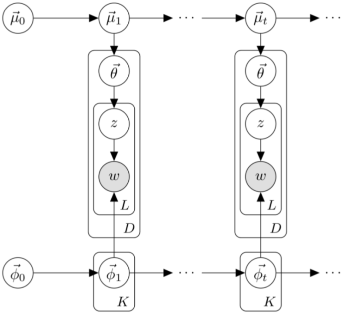
- Hierarchical Pitman-Yor Process (HPYP) for modelling discrete hierarchical priors (Teh, 2006)
- Graphical Pitman-Yor Process (GPYP) for general graphical priors (Wood and Teh, 2009)

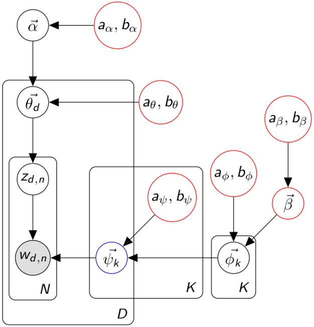
Networks of Probability Vectors

- Hierarchical Pitman-Yor Process (HPYP) for modelling discrete hierarchical priors (Teh, 2006)
- Graphical Pitman-Yor Process (GPYP) for general graphical priors (Wood and Teh, 2009)
- Block table indicator sampling for the above (Chen, Du and Buntine 2001; Du 2012)

Networks of Probability Vectors

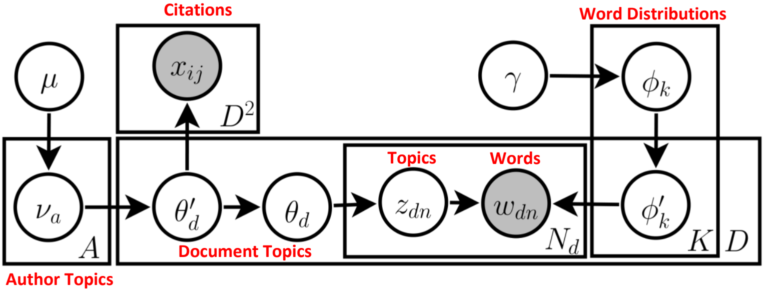
- Hierarchical Pitman-Yor Process (HPYP) for modelling discrete hierarchical priors (Teh, 2006)
- Graphical Pitman-Yor Process (GPYP) for general graphical priors (Wood and Teh, 2009)
- Block table indicator sampling for the above (Chen, Du and Buntine 2001; Du 2012)
- Fast, general modelling in complex models, *e.g.*





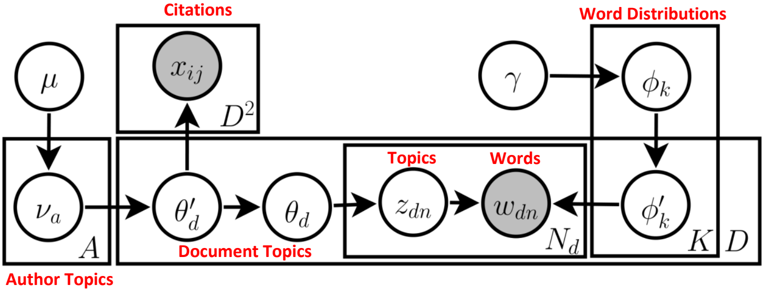
Outline

Model



- Hierarchical Pitman-Yor topic model for text.
  - Author information is captured by ν.
  - Separate topic distribution for citations and words – because they are of different types of data and citation data is given higher strength.
  - Burstiness modelling for the word distributions.

Model



$$x_{ij} \sim Poisson(\lambda_{ij}) \qquad \lambda_{ij} = \lambda_i^+ \lambda_j^- \sum_k \lambda_k^T \theta'_{ik} \theta'_{jk}$$

- Citation network Poisson model for citations.
  - λ_i^+: propensity to cite
  - λ_j^-: popularity of a document
  - λ_k^T: topic adjustment parameter
- Citations are related to the topic similarity between documents.

Posterior Inference

- Standard posterior inference procedure for topic models is work with counts rather than probability vectors.
  - Incorporating citation information naively breaks this property.
  - Our novel inference algorithm allows citation information to be treated as counts in the topic model.
  - Assumption: Connection between two documents is mainly determined by their dominant topics – reasonable in practice (see details in paper).

Posterior Inference

- Standard posterior inference procedure for topic models is work with counts rather than probability vectors.
  - Incorporating citation information naively breaks this property.
  - Our novel inference algorithm allows citation information to be treated as counts in the topic model.
  - Assumption: Connection between two documents is mainly determined by their dominant topics – reasonable in practice (see details in paper).
- Inference algorithms:
  - Collapsed Gibbs sampler for the Hierarchical PYP topic model.
    - Current state-of-the-art for Hierarchical PYP (Chen et al., 2011).
    - Note: Algorithm can be readily used without modification due to the above assumption.

Posterior Inference

- Standard posterior inference procedure for topic models is work with counts rather than probability vectors.
  - Incorporating citation information naively breaks this property.
  - Our novel inference algorithm allows citation information to be treated as counts in the topic model.
  - Assumption: Connection between two documents is mainly determined by their dominant topics – reasonable in practice (see details in paper).
- Inference algorithms:
  - Collapsed Gibbs sampler for the Hierarchical PYP topic model.
    - Current state-of-the-art for Hierarchical PYP (Chen et al., 2011).
    - Note: Algorithm can be readily used without modification due to the above assumption.
  - Metropolis-Hastings algorithm for citation network.
    - Concept similar to topic model sampler – decrement counts associated with a citation, sample a new topic for the citation and update the counts.
    - See paper for details.

Posterior Inference

- Standard posterior inference procedure for topic models is work with counts rather than probability vectors.
  - Incorporating citation information naively breaks this property.
  - Our novel inference algorithm allows citation information to be treated as counts in the topic model.
  - Assumption: Connection between two documents is mainly determined by their dominant topics – reasonable in practice (see details in paper).
- Inference algorithms:
  - Collapsed Gibbs sampler for the Hierarchical PYP topic model.
    - Current state-of-the-art for Hierarchical PYP (Chen et al., 2011).
    - Note: Algorithm can be readily used without modification due to the above assumption.
  - Metropolis-Hastings algorithm for citation network.
    - Concept similar to topic model sampler – decrement counts associated with a citation, sample a new topic for the citation and update the counts.
    - See paper for details.
  - Hyperparameter sampling to learn the hyperparameters automatically.

Outline

Data

- 3 corpus queried from CiteSeer^X and 3 existing corpus from LINQS[1].
- Keywords from Microsoft Academic Search is used to query CiteSeer^X:
  - ML: Machine Learning publications.
  - M10: Publications from Multidisciplines.
  - AvS: Arts publications *vs.* Science publications.
- From LINQS:
  - AI: Artificial Intelligence publications from CiteSeer.
  - Cora: Machine Learning publications.
  - PubMed: Publications from PubMed database on diabetes.

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

[1] Lise's INQuisitive Students, statistical relational learning group.

Experiments

- Baselines:
  - HDP-LDA with burstiness (Buntine and Mishra, 2014);
  - Non-parametric extension of author-topic model (Rosen-Zvi et al., 2004);
  - Poisson mixed-topic link model (PMTLM) (Zhu et al., 2013).

Experiments

- Baselines:
  - HDP-LDA with burstiness (Buntine and Mishra, 2014);
  - Non-parametric extension of author-topic model (Rosen-Zvi et al., 2004);
  - Poisson mixed-topic link model (PMTLM) (Zhu et al., 2013).
- Quantitative evaluations:
  - Goodness of fit test
    - Perplexity
    - Convergence
  - Document clustering
    - Purity
    - Normalised mutual information

Experiments

- Baselines:
  - HDP-LDA with burstiness (Buntine and Mishra, 2014);
  - Non-parametric extension of author-topic model (Rosen-Zvi et al., 2004);
  - Poisson mixed-topic link model (PMTLM) (Zhu et al., 2013).
- Quantitative evaluations:
  - Goodness of fit test
    - Perplexity
    - Convergence
  - Document clustering
    - Purity
    - Normalised mutual information
- Qualitative analysis:
  - Analysing topic summary from the topic-word distributions.
  - Extracting authors' interest from the author-topic distributions.
  - Graphically visualise author-topics network.

Goodness of Fit

- Perplexity:
  - Negatively related to log likelihood of the model, so lower is better.
  - Use "document completion" approach, but instead of using half of the document to estimate θ, use only the publication title to estimate θ and evaluate on the rest of the words.

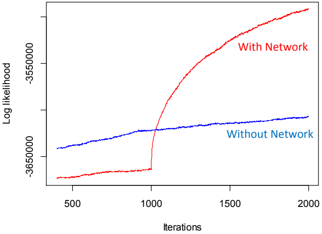
	ML		M10	
	Train	Test	Train	Test
Bursty HDP-LDA	$4904.24_{\pm 71.34}$	$4992.94_{\pm 65.57}$	$1959.36_{\pm 32.77}$	$2265.18_{\pm 68.19}$
Non-parametric ATM	$2238.19_{\pm 12.22}$	$2460.28_{\pm 11.34}$	$1562.85_{\pm 18.11}$	$1814.03_{\pm 23.18}$
CNTM w/o network	$1918.21_{\pm 4.31}$	$2057.61_{\pm 3.56}$	$912.69_{\pm 10.94}$	$1186.11_{\pm 8.32}$
CNTM w network	$\mathbf{1851.82}_{\pm 8.50}$	$\mathbf{1990.78}_{\pm 11.36}$	$\mathbf{824.04}_{\pm 11.96}$	$\mathbf{1048.33}_{\pm 21.39}$

Goodness of Fit

- Perplexity:
  - Negatively related to log likelihood of the model, so lower is better.
  - Use "document completion" approach, but instead of using half of the document to estimate θ, use only the publication title to estimate θ and evaluate on the rest of the words.

	ML		M10	
	Train	Test	Train	Test
Bursty HDP-LDA	$4904.24_{\pm 71.34}$	$4992.94_{\pm 65.57}$	$1959.36_{\pm 32.77}$	$2265.18_{\pm 68.19}$
Non-parametric ATM	$2238.19_{\pm 12.22}$	$2460.28_{\pm 11.34}$	$1562.85_{\pm 18.11}$	$1814.03_{\pm 23.18}$
CNTM w/o network	$1918.21_{\pm 4.31}$	$2057.61_{\pm 3.56}$	$912.69_{\pm 10.94}$	$1186.11_{\pm 8.32}$
CNTM w network	$\mathbf{1851.82}_{\pm 8.50}$	$\mathbf{1990.78}_{\pm 11.36}$	$\mathbf{824.04}_{\pm 11.96}$	$\mathbf{1048.33}_{\pm 21.39}$

- Convergence:
  - Plot of log likelihood over training iterations.



Qualitative Analysis

- Topic summary for the Machine Learning dataset:
  - Top words from the document-topic distributions ϕ.

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

Qualitative Analysis

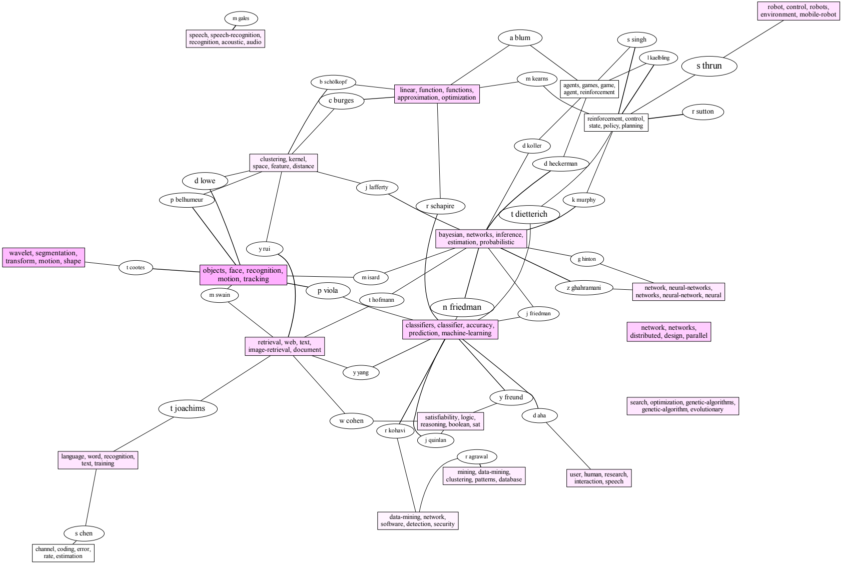
- Topic summary for the Machine Learning dataset:
  - Top words from the document-topic distributions ϕ.

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

- Authors' interest:
  - Only the dominant topic is shown.

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

Visualisation of Authors and Learned Topics



Outline

Summary

- CNTM jointly model publications' content and their citation network using auxiliary info such as authors.
- Novel sampling algorithm for the citation network allows citations to be treated like words just as in topic models.

Summary

- CNTM jointly model publications' content and their citation network using auxiliary info such as authors.
- Novel sampling algorithm for the citation network allows citations to be treated like words just as in topic models.

Thank you!