Nonparametric Bayesian Topic Modelling with Auxiliary Data

Kar Wai Lim

2 Dec 2015

Outline

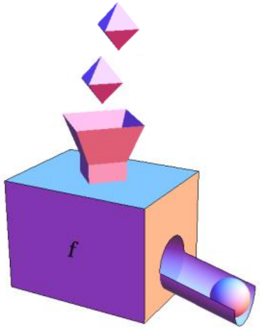
- Motivation
- Contributions
- Summary

Motivation

- Abundance of information online.

Motivation

- Abundance of information online.
- Impossible to go through them all manually.
- Need a way to process these information automatically.



Motivation

- What is a topic model?
  - A model that assign topic labels to each word.

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts" **"Budgets"** **"Children"** **"Education"**

Motivation

- What is a topic model?
  - A model that assign topic labels to each word.
  - Gives a summary of the corpus in the form of 'topics'.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Motivation

- What is a topic model?
  - A model that assign topic labels to each word.
  - Gives a summary of the corpus in the form of ‘topics’.
  - Other usages include clustering of documents.

- Examples:
  - Latent Dirichlet Allocation (LDA)
  - Author topic model (ATM)
  - HDP-LDA

Bayesian (parametric)

Bayesian (nonparametric)

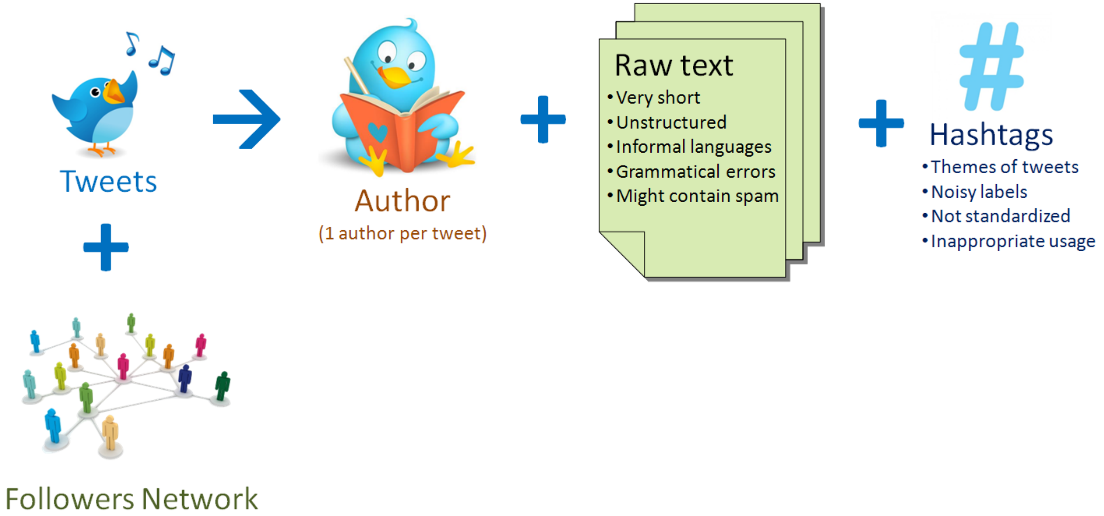
Motivation

- Why Bayesian?
  - Model-based, thus fundamentally sound compared to rule-based method.
  - Incorporation of prior information (from expert knowledge or previous experiments).
  - Clear inference!

- Why Nonparametric?
  - Very flexible.
  - Able to learn the number of clusters in topic modelling.

Motivation

- Aims:
  - Incorporate auxiliary information to improve topic modelling.



Motivation

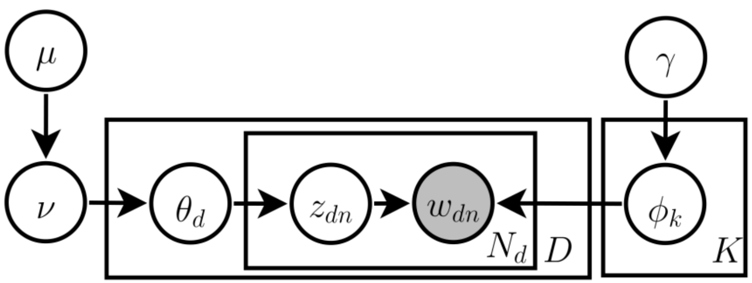
- Aims:
  - Incorporate auxiliary information to improve topic modelling.
  - Employ states-of-the-art nonparametric Bayesian techniques for NLP applications.
    - Pitman-Yor process (PYP),
    - Gaussian process (GP),
    - Inference with blocked Gibbs sampling.
  - Design a framework to implement arbitrary topic models that utilise hierarchical PYP (HPYP).
    - Speed up topic model development.

Outline

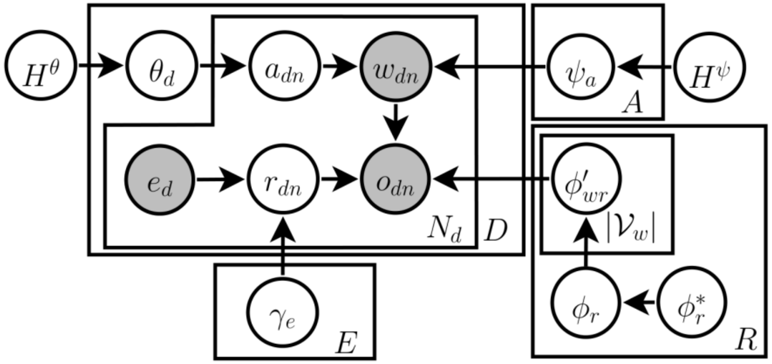
- ~~Motivation~~
- Contributions
  - Implementation Framework
  - Opinion Mining on Products
  - Bibliographic Analysis
  - Tweets Exploration
- Summary

Implementation Framework

- Want a framework to implement arbitrary HPYP topic models.



A simple HPYP topic model



A more complicated HPYP topic model

Implementation Framework

- Want a framework to implement arbitrary HPYP topic models.
  - Focus on code reusability.
  - Modularisation of the PYPs.
    - Each PYP stores variables locally.
    - Each PYP performs computation locally.
    - Each PYP can call methods of its parent PYPs, allowing recursion.
  - (We won’t go into details since they require knowledge on “Chinese Restaurant Process”)

Outline

- ~~Motivation~~
- Contributions
  - ~~Implementation Framework~~
  - Opinion Mining on Products
  - Bibliographic Analysis
  - Tweets Exploration
- Summary

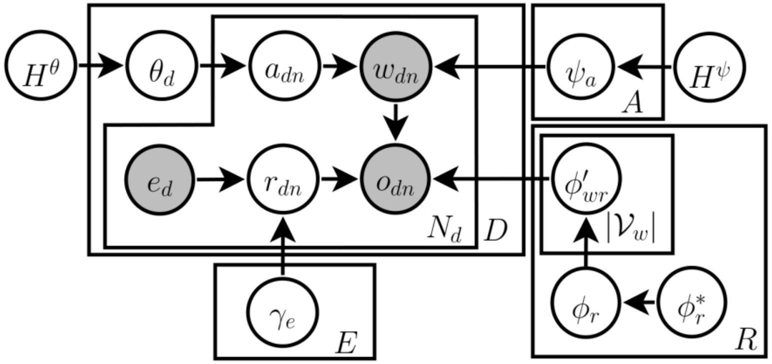
Opinion Mining

- Motivation:
  - First hand opinions on products and services are readily available on social media, i.e. tweets.
  - Good source for reviews.
  - Available in large quantity but unstructured and messy.

270 tweets on iphone 6s created last 20 mins

Twitter Opinion Topic Model

- We propose the Twitter Opinion Topic Model (TOTM) to perform opinion mining on electronic products from tweets.
- TOTM uses hashtags, mentions, emoticons and sentiment lexicons to improve sentiment analysis.



Twitter Opinion Topic Model

- Advantages of TOTM:
  - Models the target-opinion interaction directly, allowing the discovery of target-dependent opinions.
  - Example 1:
    - Long battery life is good.
    - Long working hour is bad.
  - Example 2:
    - TOTM knows that “friendly dumpling” is unlikely while existing models don’t.

A friendly dumpling

Twitter Opinion Topic Model

- Advantages of TOTM:
  - Models the target-opinion interaction directly, allowing the discovery of target-dependent opinions.
  - Provides a novel formulation to incorporate sentiment lexicon as prior into topic models.
    - Uses a tunable parameter to control the strength of the sentiment prior.
    - The tunable parameter is automatically learned and updated based on data.

Twitter Opinion Topic Model

- Advantages of TOTM:
  - Models the target-opinion interaction directly, allowing the discovery of target-dependent opinions.
  - Provides a novel formulation to incorporate sentiment lexicon as prior into topic models.
  - Enables new ways to visualise and summarise the tweet corpus.
    - Product clustering.
    - Target specific sentiment analysis.
    - Brand comparison.
    - Opinion extraction.

Data Statistics

- Main dataset:
  - Tweets containing electronic products are queried from the Twitter7 dataset.
  - This gives ~9 million tweets on electronic products.

Categories	Query Words
Mobile phones	iphone, blackberry, nokia, palmpre, sony, motorola, phone, samsung, lg, scanner, android, ios, apple
Computers	sony, dell, lenovo, toshiba, acer, asus, macbook, hp, alienware, laptop, tablet, netbook, ipad, ipod, printer, panasonic, epson, samsung, ibm, sony, microsoft, computer, windows, operatingsystem, apple
Cameras	sony, canon, nikon, camera, panasonic, epson, samsung, lg, fujitsu, kodak
Printers/ Scanners	sony, canon, nikon, dell, lenovo, toshiba, hp, printer, panasonic, epson, samsung, kyocera, lg, scanner, kodak
Gaming	xbox, playstation, wii, nintendo, gameboy, sega, squareenix

Data

- Main dataset:
  - Tweets containing electronic products are queried from the Twitter7 dataset.
  - This gives ~9 million tweets on electronic products.

- Additional datasets:
  - Sentiment140 obtained online.
    - It has 800k positive tweets and 800k negative tweets.
  - SemEval 2013 tweets.
    - 6322 manually annotated tweets.

Experiments

- Goodness-of-fit test:
  - Perplexity is commonly used to evaluate topic models.
  - Negatively related to the log likelihood of observed words, so lower perplexity is better.

Dataset	Models	Target Perplexity	Opinion Perplexity	Overall Perplexity
Electronic Product	LDA-DP	N/A	510.15 ± 0.08	N/A
	ILDA	594.81 ± 13.61	519.84 ± 0.43	556.03 ± 6.22
	TOTM	592.91 ± 13.86	**137.42** ± 0.28	**285.42** ± 3.23
Sent140	LDA-DP	N/A	329.92 ± 16.58	N/A
	ILDA	567.22 ± 16.31	306.79 ± 0.15	417.12 ± 6.12
	TOTM	530.08 ± 5.23	**93.89** ± 0.41	**223.09** ± 0.63
SemEval	LDA-DP	N/A	688.54 ± 62.17	N/A
	ILDA	2695.39 ± 65.33	433.20 ± 1.50	1080.51 ± 13.75
	TOTM	2725.51 ± 71.88	**249.04** ± 4.09	**823.74** ± 7.68

Due to modelling the target-opinion interaction directly

Experiments

- Sentiment classification:
  - Compare predicted sentiment against ground truth. (Electronic Product dataset has no sentiment labels)

Dataset	Models	Accuracy	Precision	Recall	F1-score
Sent140	LDA-DP	57.3	56.1	90.1	69.2
	ILDA	54.1	56.9	55.3	55.9
	TOTM	**65.0**	**61.7**	**90.2**	**73.3**
SemEval	LDA-DP	52.1	65.0	58.3	61.4
	ILDA	46.8	60.7	53.6	56.3
	TOTM	**73.3**	**84.0**	**74.9**	**79.0**

Qualitative Results

- Target-specific sentiment analysis.
  - Obtained by inspecting the top words in the target-sentiment-opinion distributions.

Target (w)	Sentiment (r)	Opinions (o)
phone	+1	**mobile smart** good great f***ing
	−1	**dead** damn stupid bad crazy
battery life	+1	good **long** great **7hr ultralong**
	−1	terrible poor bad horrible **non-existence**
game	+1	great good awesome favorite **cat-and-mouse**
	−1	**addictive** stupid free **full addicting**
sausage	+1	**hot grilled** good **sweet** awesome
	−1	silly **argentinian cold** huge stupid

Words in bold are target-specific opinions.

Qualitative Results

- Brand comparison:
  - The major brands are from hashtags and mentions.

Brands	Sentiment	Aspects / Targets' Opinions: Camera	Aspects / Targets' Opinions: Phone
Canon	+1	*camera* → great compact amazing *pictures* → great nice creative	
	−1	*camera* → expensive small bad *lens* → prime cheap broken	
Sony	+1	*photos* → great lovely amazing *camera* → good great nice	*phone* → great smart beautiful *reception* → perfect
	−1	*camera* → big crappy defective *lens* → vertical cheap wide	*phone* → worst crappy shittest *battery life* → low
Samsung	+1	*camera* → gorgeous great cool *pics* → nice great perfect	*phone* → mobile great nice *service* → good sweet friendly
	−1	*camera* → digital free crazy *shots* → quick wide	*phone* → stupid bad fake *battery life* → solid poor terrible

Reference

- Refer to the paper

    Kar Wai Lim and Wray Buntine. 2014. Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (CIKM '14). ACM.

  for details.

Outline

- ~~Motivation~~
- Contributions
  - ~~Implementation Framework~~
  - ~~Opinion Mining on Products~~
  - Bibliographic Analysis
  - Tweets Exploration
- Summary

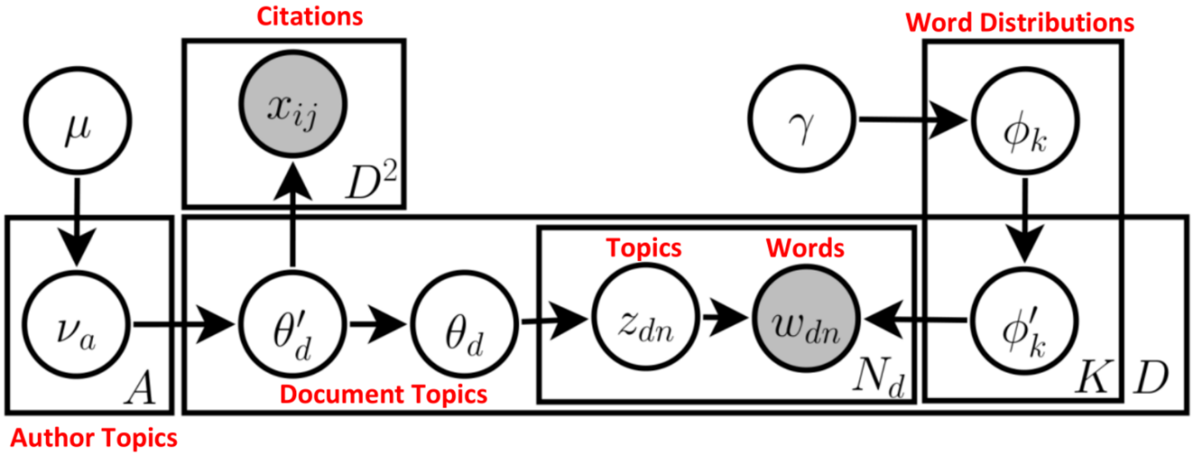
Bibliographic Analysis

- Motivation:
  - Research publications are readily available.
  - They are accompanied by auxiliary information such as authors, categories, publishers etc.
  - Another interesting information is the citation network.

- Aim:
  - Utilise these auxiliary data for bibliographic analysis on research publications.

Citation Network Topic Model

- We propose the Citation Network Topic Model (CNTM), comprised of:
  - A HPYP topic model (an extension of the ATM) that models text and incorporates authorship information.
  - Citation network Poisson model for the citations.



Citation Network Topic Model

- We propose the Citation Network Topic Model (CNTM), comprised of:
  - A HPYP topic model (an extension of the ATM) that models text and incorporates authorship information.
  - Citation network Poisson model for the citations.

- We also propose a method to incorporate supervision into topic modelling.
  - Using categorical information.

Citation Network Topic Model

- Novelty in posterior inference:
  - Standard posterior inference procedure for topic models is with counts rather than probability vectors.
  - Incorporating citation information naively breaks this property.
  - We propose a novel inference algorithm that 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.

Data

- Datasets:
  - Query 3 datasets from CiteSeerX.
  - Additional 3 datasets from LINQS.

Dataset	Classes	Categorical Labels
ML	1	Machine Learning
M10	10	Agriculture, Archaeology, Biology, Computer Science, Physics, Financial Economics, Industrial Engineering, Material Science, Petroleum Chemistry, Social Science
AvS	5	History, Religion, Physics, Chemistry, Biology
CS	6	Agents, AI, DB, IR, ML, HCI
Cora	7	Case Based, Genetic Algorithms, Neural Networks, Theory, Probabilistic Methods, Reinforcement Learning, Rule Learning
PubMed	3	"Diabetes Mellitus, Experimental", Diabetes Mellitus Type 1, Diabetes Mellitus Type 2

Experiments

- Goodness-of-fit test:

Model	Perplexity Train	Test	Train	Test
	ML		**M10**	
Bursty HDP-LDA	4904.2 ± 71.3	4992.9 ± 65.6	2467.9 ± 34.8	2825.6 ± 61.4
Non-parametric ATM	2238.2 ± 12.2	2460.3 ± 11.3	1822.4 ± 15.0	2056.4 ± 18.3
CNTM w/o network	2036.3 ± 4.6	2118.1 ± 3.7	922.6 ± 11.0	1263.9 ± 8.8
CNTM w network	**1919.5** ± 8.8	**2039.5** ± 11.7	**910.2** ± 13.3	**1261.0** ± 25.7
	AvS		**CS**	
Bursty HDP-LDA	2460.4 ± 66.4	2612.8 ± 91.7	1498.4 ± 4.1	1616.8 ± 38.8
Non-parametric ATM	2225.9 ± 45.5	2511.9 ± 52.4	N/A	N/A
CNTM w/o network	1540.2 ± 18.5	1959.2 ± 2.4	1506.8 ± 4.4	1609.5 ± 39.2
CNTM w network	**1515.9** ± 2.1	**1938.9** ± 10.4	**1168.6** ± 27.3	**1588.2** ± 93.9
	Cora		**PubMed**	
Bursty HDP-LDA	678.3 ± 1.7	**706.3** ± 16.8	300.0 ± 0.3	300.2 ± 1.2
CNTM w/o network	554.8 ± 14.1	881.1 ± 110.9	**299.9** ± 0.2	300.1 ± 1.3
CNTM w network	**527.0** ± 8.7	719.0 ± 111.4	350.5 ± 20.1	**297.3** ± 3.2

Experiments

- Document clustering:
  - Compare documents grouped by topics with ground truth categorical labels.

Dataset	Model	Purity			NMI		
		Train	Test	Overall	Train	Test	Overall
M10	Bursty HDP-LDA	61.7	65.6	62.1	34.8	67.0	38.0
	Non-parametric ATM	55.4	57.8	55.7	29.1	63.0	32.4
	CNTM w/o network	67.3	64.9	67.0	42.5	66.5	44.9
	CNTM w network	66.4	**69.9**	66.8	41.1	**68.6**	43.8
	SCNTM ($\eta = 10$)	85.3	53.1	82.1	60.4	62.7	60.6
	SCNTM ($\eta = \infty$)	**88.1**	47.8	**84.0**	**62.3**	62.3	**62.3**
AvS	Bursty HDP-LDA	72.8	75.0	73.0	32.1	66.3	35.5
	Non-parametric ATM	64.1	65.2	64.2	24.7	61.9	28.4
	CNTM w/o network	77.0	**76.3**	76.9	37.4	66.6	40.3
	CNTM w network	76.0	74.0	75.8	35.4	65.5	38.4
	SCNTM ($\eta = 10$)	**87.9**	67.3	**85.8**	47.5	**66.7**	**49.4**
	SCNTM ($\eta = \infty$)	87.1	50.5	83.4	**47.8**	64.5	**49.4**

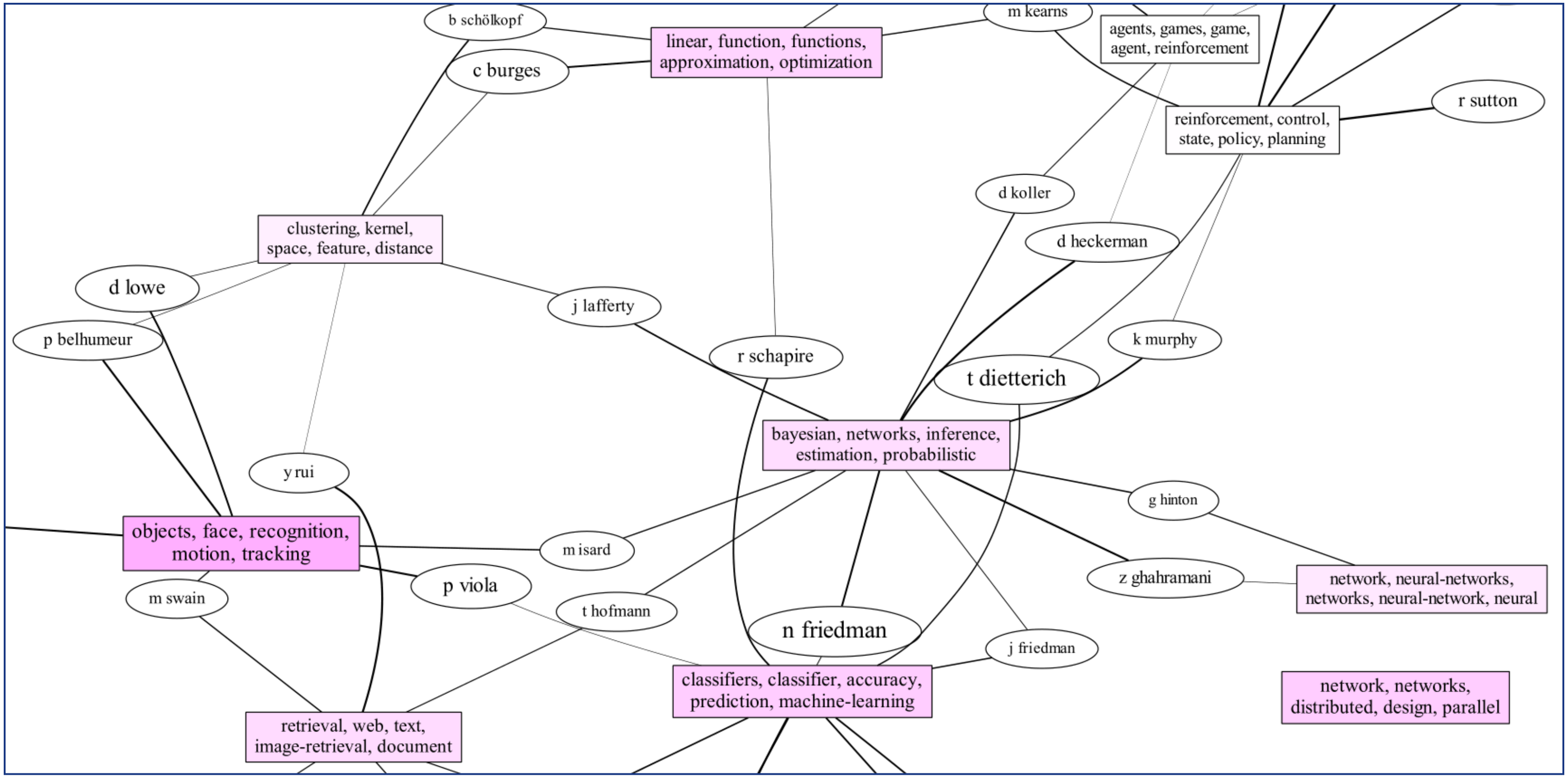
Qualitative Results

- Topical summary extraction:

Topic	Top Words
	ML
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
	M10
DNA Sequencing	genes, gene, sequence, binding sites, dna
Agriculture	soil, water, content, soils, ground
Financial Market	volatility, market, models, risk, price
Bayesian Modelling	bayesian, methods, models, probabilistic, estimation
Quantum Theory	quantum, theory, quantum mechanics, classical, quantum field
	AvS
Language Modelling	type, polymorphism, types, language, systems
Molecular Structure	copper, protein, model, water, structure
Quantum Theory	theory, quantum, model, quantum mechanics, systems
Social Science	research, development, countries, information, south africa
Family Well-being	children, health, research, social, women

Qualitative Results

- Author-topics network visualisation:



Reference

- Refer to the paper

  Kar Wai Lim and Wray Buntine. 2014. Bibliographic analysis with the Citation Network Topic Model. In Proceedings of the Sixth Asian Conference on Machine Learning (ACML ‘14).

  for details.

Outline

- ~~Motivation~~
- Contributions
  - ~~Implementation Framework~~
  - ~~Opinion Mining on Products~~
  - ~~Bibliographic Analysis~~
  - Tweets Exploration
- Summary

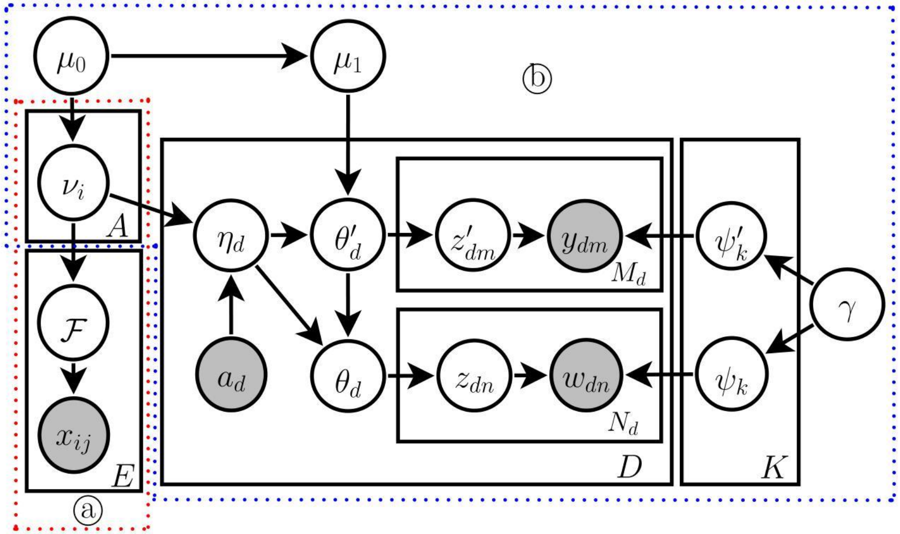
Tweets Exploration

- Motivation:
  - Tweets are short, unstructured, often contain errors.
  - Tweets are informal – laden with user-defined abbreviations and hashtags.
  - Vanilla topic models do not well work on tweets.

- Aim:
  - Make use of available information and design a topic model that works well on tweets.

Twitter Network Topic Model

- We propose the Twitter Network Topic Model (TNTM), which uses:
  a) A GP network model that models the followers network.
  b) A HPYP topic model that models authors, text and hashtags (the hashtags are treated as words).



Data

- Datasets:
  - T6 dataset is queried from the Twitter7 dataset using keywords.
  - The other 3 datasets are from Mehrotra et al. (2013).

Dataset	Queries
T6	#sport, #music, #finance, #politics, #science and #tech
Generic	business, design, family, food, fun, health, movie, music, space, sport
Specific	Apple, baseball, Burgerking, cricket, France, Mcdonalds, Microsoft, Obama, Sarkozy, United States
Events	attack, conference, Flight 447, Iran election, Jackson, Lakers, recession, scandal, swine flu, T20

Experiments

- Ablation test:
  - Comparison of the full TNTM model with ablated counterparts (with some component removed).

TNTM Model	Test Perplexity	Network Log Likelihood
No author	669.12 $\pm$ 9.3	N/A
No hashtag	1017.23 $\pm$ 27.5	-522.83 ± 17.7
No μ_1 node	607.70 $\pm$ 10.7	-508.59 ± 9.8
No θ'-θ connection	551.78 $\pm$ 16.0	-509.21 ± 18.7
No power-law	508.64 $\pm$ 7.1	-560.28 ± 30.7
Full model	**505.01** $\pm$ 7.8	$-\mathbf{500.63} \pm 13.6$

Experiments

- Document clustering:
  - Compared against LDA with different pooling schemes.

Method/Model	Purity			NMI		
	Generic	**Specific**	**Events**	**Generic**	**Specific**	**Events**
No pooling	0.49	0.64	0.69	0.28	0.22	0.39
Author	0.54	0.62	0.60	0.24	0.17	0.41
Hourly	0.45	0.61	0.61	0.07	0.09	0.32
Burstwise	0.42	0.60	0.64	0.18	0.16	0.33
Hashtag	0.54	**0.68**	0.71	0.28	0.23	0.42
TNTM	**0.66**	**0.68**	**0.79**	**0.43**	**0.31**	**0.52**

Qualitative Results

- Topic labelling using hashtags:
  - Hashtags are good candidates for topic labels.

Topic	Top Hashtags	Top Words
Topic 1	finance, money, economy	finance, money, bank, marketwatch, stocks, china, group, shares, sales
Topic 2	politics, iranelection, tcot	politics, iran, iranelection, tcot, tlot, topprog, obama, musiceanewsfeed
Topic 3	music, folk, pop	music, folk, monster, head, pop, free, indie, album, gratuit, dernier
Topic 4	sports, women, asheville	sports, women, football, win, game, top, world, asheville, vols, team
Topic 5	tech, news, jobs	tech, news, jquery, jobs, hiring, gizmos, google, reuters
Topic 6	science, news, biology	science, news, source, study, scientists, cancer, researchers, brain, biology, health

Qualitative Results

- Authors inspection:
  - We look at several active tweeters and their dominant topic.

Author	Dominant Topic
finance_yard	finance, money, realestate
ultimate_music	music, ultimatemusiclist, mp3
seriouslytech	technology, web, tech
seriouspolitics	politics, postrank, news
pr_science	science, news, postrank

Reference

- Refer to the paper

  Kar Wai Lim, Changyou Chen and Wray Buntine. 2013. Twitter-Network Topic Model: A full Bayesian treatment for social network and text modeling. In Advances in Neural Information Processing Systems: Topic Models Workshop. NIPS Workshop 2013.

  for details.

Outline

- ~~Motivation~~
- ~~Contributions~~
  - ~~Implementation Framework~~
  - ~~Opinion Mining on Products~~
  - ~~Bibliographic Analysis~~
  - ~~Tweets Exploration~~
- Summary

Summary

- We proposed 3 nonparametric Bayesian topic models that incorporate auxiliary information for NLP tasks:
  - Opinion mining and sentiment analysis.
  - Bibliographic analysis.
  - Text data exploration.
- We provided a framework to implement these topic models.
- We found that topic models fit better to the data as we utilise more auxiliary information.