Scaling the Dynamic Author-Persona Topic Model to Billion Word Corpora

Robert Giaquinto and Arindam Banerjee, University of Minnesota
{giaquinto.ra@gmail.com, banerjee42@gmail.com}

Minnesota Supercomputing Institute
University of Minnesota
Driven to Discover®

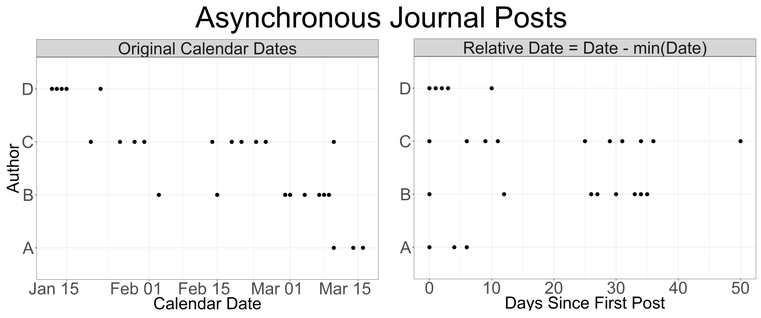
Introduction

- **Objective**: Design a new topic model to meet challenges presented by the CaringBridge (CB) dataset.
  - CB dataset: patients and caregivers writing during a health crisis.
- **Challenges:** Asynchronous nature of CB journals, massive collection of texts.
- **Contributions:**
  - Design **Dynamic Author-Persona (DAP)** topic model for multiple authors writing over time [6].
    - Represents authors by a persona – personas cluster authors with similar topic trajectories.
  - Introduce **regularized variational inference (RVI)** algorithm to encourage personas to be distinct.
  - Adapt new ideas in approximate inference to DAP, leading to **DAP Performed Exceedingly Rapidly (DAPPER)** [7,8].
- **Results:**
  - Better likelihood's compared to competing models.
  - Demonstrate scalability on CB dataset.
  - Compelling qualitative results: health journeys of CB authors.

Challenges

- **Asynchronous journals.**
  - Authors start and stop journaling at different dates and states of health.
  - Long lasting chronic condition vs. brief ailment.
- **Existing methods not adequate**. State-of-the-art topic models [1, 2, 3, 4, 5]:
  - Identify topics and how they change over time, or associate authors with certain topics.
  - What about common narratives and the authors sharing them (e.g. “health journeys”)?
- **Scalability of complex topic models like DAP.**
  - Non-conjugate model terms are a severe bottleneck.
  - Iterative optimizations required for each document during training.

Asynchronous Journal Posts



Data and Evaluation

- **CaringBridge Dataset:** Journals written by patients and caregivers during a health crisis.
  - 9 million journals (40GB), ~200k authors, written between 2006 – 2016.
- **Preprocessing:** 1st year of posting, <10 words / post.
- **Quantitative Evaluation:** 2,000 authors were randomly selected (114,532 journals, 90% for train, 10% for test).
  (1) Compare likelihoods for DAP and DAPPER to similar models (LDA [1], DTM [3], CDTM [4]).
  (2) Compare rate of converge of DAP and DAPPER.
- **Qualitative Evaluation:** Train DAPPER on full dataset.
  - Demonstrate quality of model output at scale.

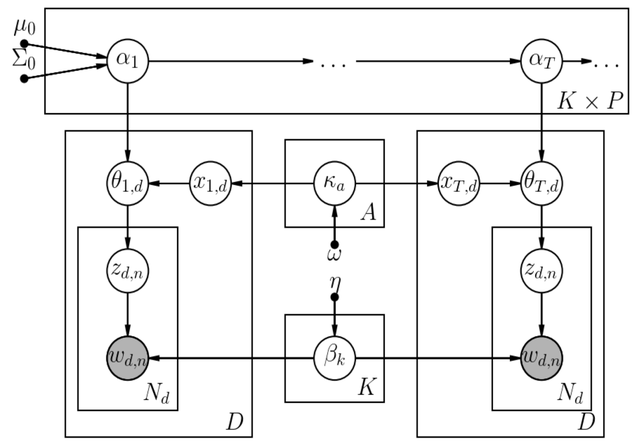
Acknowledgements

We thank University of Minnesota Supercomputing Institute, and CaringBridge for their support and collaboration. Research supported by NSF grants IIS-1563950, IIS-1447566, IIS-1447574, IIS-1422557, CCF-1451986, CNS-1314560.

References

[1] Blei, D. M., Ng, A. Y., and Jordan, M. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(4-5):993–1022.
[2] Lafferty, J. D., and Blei, D. M. 2006. Correlated Topic Models. *Advances in Neural Information Processing Systems 18* 147–154.
[3] Blei, D. M., and Lafferty, J. D. 2006. Dynamic Topic Models. *International Conference on Machine Learning* 113–120.
[4] Wang, C., Blei, D., and Heckerman, D. 2008. Continuous Time Dynamic Topic Models. *Proc of UAI* 579–586.
[5] Mimno, D., and McCallum, A. 2007. Expertise modeling for matching papers with reviewers. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* 500–509.
[6] Giaquinto, R., and Banerjee, A. 2018. Topic Modeling on Health Journals with Regularized Variational Inference. *AAAI*.
[7] Khan, M.E., and Lin, W. 2017. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. *AIStats 54*, 878–887.
[8] Giaquinto, R., and Banerjee, A. 2018. DAPPER: Scaling Dynamic Author Persona Topic Model to Billion Word Corpora. Submitted to KDD.

Dynamic Author-Persona Topic Model



Variable	Description
P	Number of personas
K	Number of topics
A	Number of authors
T	Number of time steps
D	Number of documents
N_d	Number of words in document d
$\mathbf{w}_{t,d}$	*Observed* words in document d_t
$\mathbf{z}_n$	Assigns word n to a topic
$\boldsymbol{\theta}_{t,d}$	Distribution over topics for document d_t
$\mathbf{x}_{t,d}$	Assigns author of document d_t to a persona
$\boldsymbol{\kappa}_a$	Author a's distribution over personas
$\boldsymbol{\beta}_k$	Topic k's distribution over words
$\boldsymbol{\alpha}_{t,p}$	Distribution over topics for persona p at time t

Regularized Variational Inference

- **Background:** Variational inference (VI) frames training of topic model as optimization problem.
  - Faster than classical approaches like MCMC.
- **Approach**: Add regularization *r(q)* to VI's objective function.
- **Novelty:** RVI nudges model to seek *certain* solutions.
- **Applied to DAP:** Encourage distinct personas with *r(q):*

$$r(q) \propto \sum_{q \neq p} \boldsymbol{\alpha}_p^\top \boldsymbol{\alpha}_q$$

  - Inner product between personas (excluding a persona with itself).
  - Large penalty when two personas are similar.

DAP Performed Exceedingly Rapidly

- **Approach:** Adapt Conjugate-computation variational inference (CVI) to DAP's training bottlenecks [7].
  - CVI transforms inference in non-conjugate model to a conjugate model.
  - Replaces iterative optimizations with fast, closed form updates.
- **Outcome:** Parallelizable Expectation-Maximization style of algorithm.
  - DAPPER can train on mini-batches of documents, similar to Stochastic Variational Inference.

Results

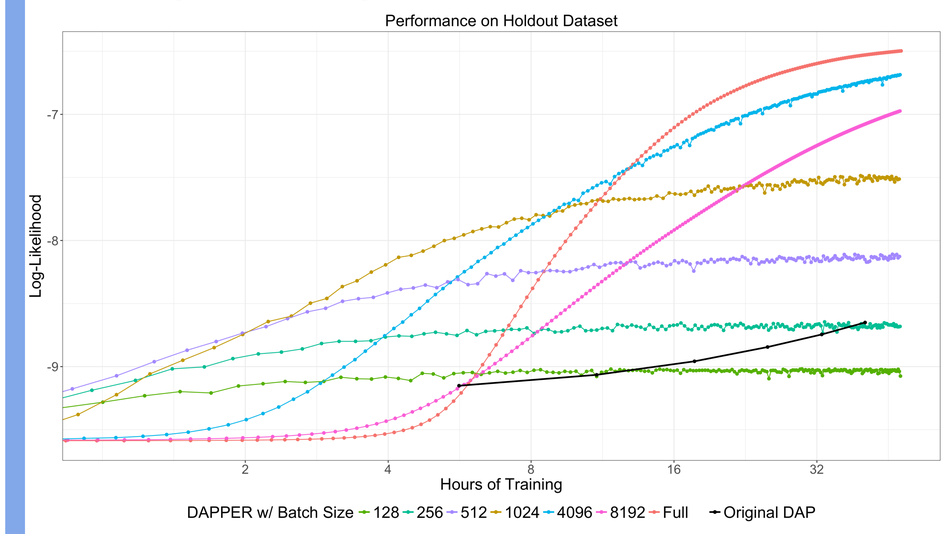
- **Quantitative (1):** Compare model performances.
  - Compute per-word log-likelihoods (PWLL) of documents in test set (*larger values are better*).

Model	PWLL, 24 Hrs, 1 CPU	PWLL Final
DAPPER (full batch)	**-6.73**	**-6.49**
DAPPER (batch size = 512)	-8.19	-8.13
DAP (ρ=0.2)	-8.09	-6.47
DAP (ρ=0.0)	-8.82	-7.22
CDTM	-8.81	-8.81
DTM	-9.59	-9.59
LDA	-9.23	-9.23

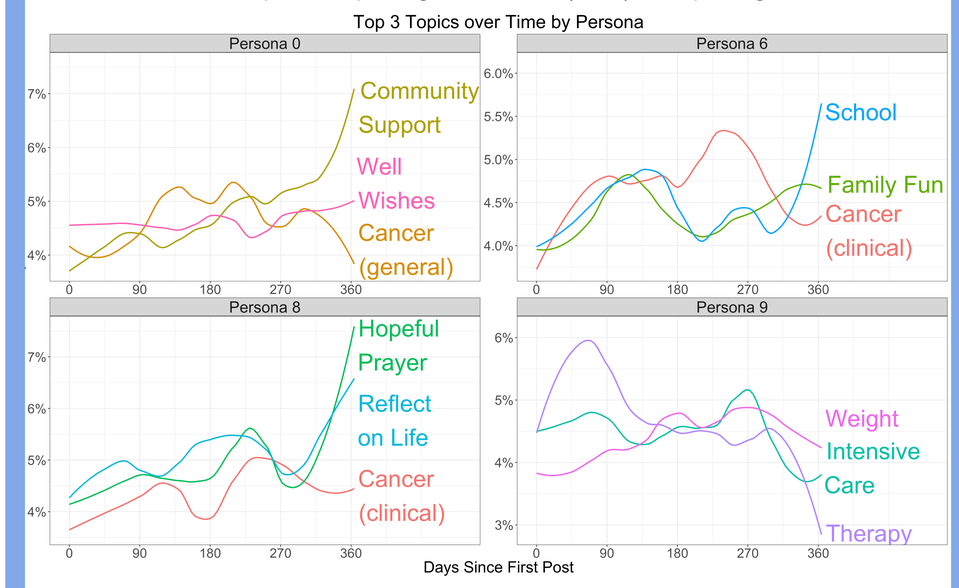
  - DAP doesn't converge in <24 hours, but still achieves competitive performance.
  - DAP trained with RVI (rho=0.2) boosts performance.
  - DAPPER achieves state-of-the-art performance with using full gradients (no mini-batches).

Results

- **Quantitative (2):** DAPPER significantly faster than predecessor.
  - Each DAP epoch takes ~6 hours; DAPPER: <10 minutes.
  - Small batch sizes converge quickly.
  - Large batch sizes generalize better on test data.



- **Qualitative:** Scale DAPPER to full 9 million journal dataset using Minnesota Supercomputing Institute's (MSI) computing resources.



- Training on parallel nodes leads to convergence in <48 hours.
- Personas are distinct, capture coherent health journeys.
- Personas 0 engages with community, and less clinical when writing about cancer.
- Personas 6 and 8 write about cancer using clinical terminology.
  - Persona 6's non-health updates on school, family, celebrations.
  - Persona 8's non-health updates are deep, reflective, prayerful.

Community Support	Physical Therapy	Reflect on Life	Hopeful Prayer	Family Fun	Infection	Weather	School
family	therapy	life	god	christmas	blood	nice	school
friend	rehab	know	pray	play	infection	weather	shot
church	therapist	child	prayer	birthday	fluid	walk	go
thank	physical	never	lord	game	fever	lunch	appt
card	pt	love	bless	fun	antibiotic	cold	class
love	chair	year	please	kid	pressure	snow	tomorrow
service	speech	live	heal	party	kidney	outside	grandma
friends	progress	people	trust	year	iv	breakfast	teacher
support	move	cancer	peace	enjoy	lung	rain	home
gift	arm	moment	continue	dinner	clot	go	aunt

Cancer (clinical)	Cancer (general)	Intensive Care	Well Wishes	Hair Loss	Surgery	Bedtime	Weight
chemo	cancer	tube	dad	hair	surgery	sleep	weight
blood	treatment	breathe	mom	leg	surgeon	night	mommy
count	radiation	oxygen	everyone	wear	heart	bed	gain
bone	scan	lung	message	head	dr	wake	feed
marrow	chemo	feed	guestbook	look	office	nurse	daddy
platelet	tumor	x_ray	please	cut	op	say	bottle
round	oncologist	chest	prayer	knee	procedure	asleep	pound
clinic	dr	nurse	read	hat	cardiologist	_time_	feeding
transfusion	ct	vent	visit	wig	valve	room	oz
url	result	stomach	update	shave	ha	tell	milk

Table 2: Top 10 words associated with the most prevalent topics found by the DAP model ($\rho = 0.2$). Topic labels are selected manually in order to aid reference with Figure 3. The words _time_ and _URL_ refer to the result of text pre-processing steps for capturing common patterns like the time of day and website URLs, respectively.

Conclusions

- DAP is uniquely suited to model text data with a temporal structure and written by multiple authors.
- DAP discovers latent personas – a novel component that identifies authors with similar topics trajectories.
- RVI algorithm improves performance.
- Adapting CVI to DAP leads to DAPPER (35x faster and better performance).
- DAPPER scales to massive datasets, like CB journals, and takes advantage of parallel computing resources.