Twitter-Network Topic Model: A Full Bayesian Treatment for Social Network and Text Modeling

Kar Wai Lim
ANU, NICTA
Canberra, Australia

Changyou Chen
ANU, NICTA
Canberra, Australia

Wray Buntine
NICTA, ANU
Canberra, Australia

Abstract

Twitter data is extremely noisy – each tweet is short, unstructured and with informal language, a challenge for current topic modeling. On the other hand, tweets are accompanied by extra information such as authorship, hashtags and the user-follower network. Exploiting this additional information, we propose the *Twitter-Network* (TN) topic model to jointly model the text and the social network in a full Bayesian nonparametric way. The TN topic model employs the hierarchical Poisson-Dirichlet processes (PDP) for text modeling and a Gaussian process random function model for social network modeling. We show that the TN topic model significantly outperforms several existing nonparametric models due to its flexibility. Moreover, the TN topic model enables additional informative inference such as authors' interests, hashtag analysis, as well as leading to further applications such as author recommendation, automatic topic labeling and hashtag suggestion. Note our general inference framework can readily be applied to other topic models with embedded PDP nodes.

1 Introduction

Emergence of web services such as blog, microblog and social networking websites allows people to contribute information publicly. This user-generated information is generally more personal, informal and often contains personal opinions. In aggregate, it can be useful for reputation analysis of entities and products, natural disasters detection, obtaining first-hand news, or even demographic analysis. Twitter, an easily accessible source of information, allows users to voice their opinions and thoughts in short text known as *tweets*.

Latent Dirichlet allocation (LDA) [1] is a popular form of topic model. Unfortunately, a direct application of LDA on tweets yields poor result as tweets are short and often noisy [2], *i.e.* tweets are unstructured and often contain grammatical and spelling errors, as well as *informal* words such as user-defined abbreviations due to the 140 characters limit. LDA fails on short tweets since it is heavily dependent on word co-occurrence. Also notable is that text in tweets may contain special tokens known as *hashtags*; they are used as keywords and allow users to link their tweets with other tweets tagged with the same hashtag. Nevertheless, hashtags are informal since they have no standards. Hashtags can be used as both inline words or categorical labels. Hence instead of being hard labels, hashtags are best treated as special words which can be the themes of the tweets. Tweets are thus challenging for topic models, and *ad hoc* alternatives are used instead. In other text analysis applications, tweets are often 'cleansed' by NLP methods such as lexical normalization [3]. However, the use of normalization is also criticized [4].

In this paper, we propose a novel method for short text modeling by leveraging the auxiliary information that accompanies tweets. This information, complementing word co-occurrence, allows us to model the tweets better, as well as opening the door to more applications, such as user recommendation and hashtag suggestion. Our main contributions include: 1) a fully Bayesian nonparametric

model called *Twitter-Network* (TN) *topic model* that models tweets very well; and 2) a combination of both the *hierarchical Poisson Dirichlet process* (HPDP) and the *Gaussian process* (GP) to jointly model text, hashtags, authors and the followers network. We also develop a flexible framework for arbitrary PDP networks, which allows quick deployment (including inference) of new variants of HPDP topic models. Despite the complexity of the TN topic model, its implementation is made relatively straightforward with the use of the framework.

2 Background and Related Work

LDA is often extended for different types of data, some notable examples that use auxiliary information are the *author-topic model* [5], the *tag-topic model* [6], and *Topic-Link LDA* [7]. However, these models only deal with just one kind of additional information and do not work well with tweets since they are designed for other types of text data. Note that the tag-topic model treats tags as hard labels and uses them to group text documents, which is not appropriate for tweets due to the noisy nature of hashtags. *Twitter-LDA* [2] and the *behavior-topic model* [8] were designed to explicitly model tweets. Both models are not admixture models since they limit one topic per document. The behavior-topic model analyzes tweets' "posting behavior" of each topic for user recommendation. On the other hand, the *biterm topic model* [9] uses only the biterm co-occurrence to model tweets, discarding document level information. Both biterm topic model and Twitter-LDA do not incorporate any auxiliary information. All the above topic models also have a limitation in that the number of topics need to be chosen in advance, which is difficult since this number is not known.

To sidestep the need of choosing the number of topics, [10] proposed *Hierarchical Dirichlet process* (HDP) LDA, which utilizes the Dirichlet process (DP) as nonparametric prior. Furthermore, one can replace the DP with the Poisson-Dirichlet process (PDP, also known as the Pitman-Yor process), which models the power-law of word frequencies distributions in natural languages. In natural languages, the distribution of word frequencies exhibits a power-law [11]. For topic models, replacing the Dirichlet distribution with the PDP can yield great improvement [12].

Some recent work models text data with network information ([7, 13, 14]), however, these models are parametric in nature and can be restrictive. On the contrary, Miller *et al.* [15] and Lloyd *et al.* [16] model network data directly with nonparametric priors, *i.e.* with the Indian Buffet process and the Gaussian process respectively, but do not model text.

3 Model Summary

The TN topic model makes use of the accompanying *hashtags*, *authors*, and *followers network* to model tweets better. The TN topic model is composed of two main components: a HPDP topic model for the text and hashtags, and a GP based random function model for the followers network. The authorship information serves to connect the two together.

We design our HPDP topic model for text as follows. First, generate the global topic distribution μ_0 that serves as a prior. Then generate the respective authors' topic distributions ν for each author, and a miscellaneous topic distribution μ_1 to capture topics that deviate from the authors' usual topics.

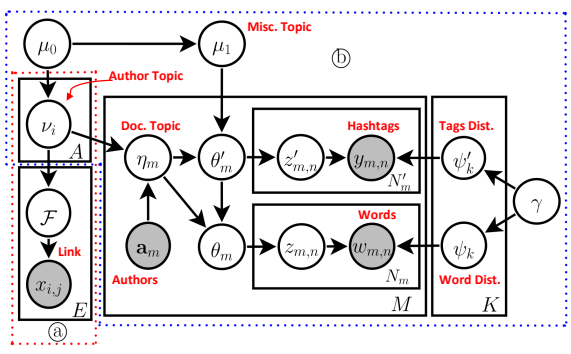

Figure 1: Twitter-Network topic model

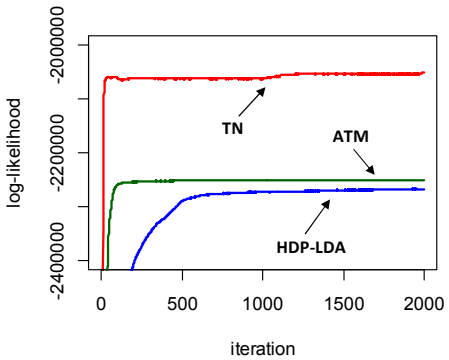

Figure 2: Log-likelihood vs. iterations

Given ν and μ_1, we generate the topic distributions for the documents, and words (η, θ', θ). We also explicitly model the influence of hashtags to words. Hashtag and word generation follows standard LDA and is not discussed here. Note that the tokens of hashtags are shared with the words, *i.e.* the hashtag *#happy* share the same token as the word *happy*. Also note that all distributions on probability vectors are modeled by the PDP, making the model a network of PDP nodes.

The network modeling is connected to the HPDP topic model *via* the author topic distributions ν, where we treat ν as inputs to the GP in the network model. The GP, denoted as $\mathcal{F}$, determines the links between the authors (x). Figure 1 displays the graphical model of TN, where region ⓐ and ⓑ shows the network model and topic model respectively. See supplementary material[1] for a detailed description. We emphasize that our treatment of the network model is different to that of [16]. We define a new kernel function based on the cosine similarity in our network model, which provides significant improvement over the original kernel function. Also, we derive a new sampling procedure for inference due to the additive coupling of topic distributions and network connections.

4 Posterior Inference

We alternatively perform Markov chain Monte Carlo (MCMC) sampling on the topic model and the network model, conditioned on each other. We derive a collapsed Gibbs sampler for the topic model, and a Metropolis-Hastings (MH) algorithm for the network model. We develop a framework to perform collapse Gibbs sampling generally on any Bayesian network of PDPs, built upon the work of [17, 18], which allows quick prototyping and development of new variants of topic model. We refer the readers to the supplementary materials for the technical details.

5 Experiments and Applications

We evaluate the TN topic model quantitatively with standard topic model measures such as test-set perplexity, likelihood convergence and clustering measures. Qualitatively, we evaluate the model by visualizing the topic summaries, authors' topic distributions and by performing an automatic labeling task. We compare our model with HDP-LDA, a nonparametric variant of the author-topic model (ATM), and the original random function network model. We also perform ablation studies to show the importance of each component in the model. The results of the comparison and ablation studies are shown in Table 1. We use two tweets corpus for experiments, first is a subset of Twitter7 dataset[2] [19], obtained by querying with certain keywords (*e.g.* finance, sports, politics). we remove tweets that are not English with *langid.py* [20] and filter authors who do not have network information and who authored less than 100 tweets. The corpus consists of 60370 tweets by 94 authors. We then randomly select 90% of the dataset as training documents and use the rest for testing. Second tweets corpus is obtained from [21], which contains a total of 781186 tweets. We note that we perform no word normalization to prevent any loss of meaning of the noisy text.

Experiment Settings In all cases, we vary α from 0.3 to 0.7 on topic nodes (μ_0, μ_1, ν_i, η_m, θ'_m, θ_m) and set $\alpha = 0.7$ on vocabulary nodes (ψ, γ) to induce power-law. We initialize β to 0.5, and set its hyperprior to Gamma$(0.1, 0.1)$. We fix the hyperparameters λ's, s, l and σ to 1 since their values have no significant impact on model performance. In the following evaluations, we run the sampling algorithms for 2000 iterations for the training likelihood to converge. We repeat each experiment five times to reduce the estimation error of the evaluation measures. In the experiments for the TN topic model, we achieve a better computational efficiency by first running the collapsed Gibbs sampling for 1000 iterations before the full inference procedure. In Figure 2, we can see that the TN topic model converges quickly compared to the HDP-LDA and the nonparametric ATM. Also, the training likelihood of the TN topic model becomes better sampling for the network information after 1000 iterations.

Automatic Topic Labeling There have been recent attempts to label topics automatically in topic modeling. Here, we show that using hashtag information allows us to get good labels for topics. Table 2 shows topics labeled by the TN topic model. More detailed topic summaries are shown in

[1] Supplementary material is available online at the authors' websites.

[2] http://snap.stanford.edu/data/twitter7.html

Table 1: Perplexity & network log-likelihood

	Perplexity	Network
HDP-LDA	$358.1_{\pm 6.7}$	N/A
ATM	$302.9_{\pm 8.1}$	N/A
Random Function	N/A	$-294.6_{\pm 5.9}$
No Author	$243.8_{\pm 3.4}$	N/A
No Hashtag	$307.5_{\pm 8.3}$	$-269.2_{\pm 9.5}$
No μ_1 node	$221.3_{\pm 3.9}$	$-271.2_{\pm 5.2}$
No Word-tag link	$217.6_{\pm 6.3}$	$-275.0_{\pm 10.1}$
No Power-law	$222.5_{\pm 3.1}$	$-280.8_{\pm 15.4}$
No Network	$218.4_{\pm 4.0}$	N/A
Full TN	$\mathbf{208.4}_{\pm 3.2}$	$\mathbf{-266.0}_{\pm 6.9}$

Table 2: Labeling topics with hashtags

	Top hashtags/words
T0	**#finance #money #economy**
	finance money bank marketwatch stocks china group
T1	**#politics #iranelection #tcot**
	politics iran iranelection tcot tlot topprog obama
T2	**#music #folk #pop**
	music folk monster head pop free indie album gratuit

Table 3: Topics by authors

Twitter ID	Top topics represented by hashtags
finance_yard	#finance #money #realestate
ultimate_music	#music #ultimatemusiclist #mp3
seriouslytech	#technology #web #tech
seriouspolitics	#politics #postrank #news
pr_science	#science #news #postrank

Table 4: Cosine similarity

Recommended	1st	2nd	3rd
Original	0.00	0.05	0.06
TN	**0.78**	**0.57**	**0.55**
Not-recommended	1st	2nd	3rd
Original	0.36	0.33	0.14
TN	**0.17**	**0.09**	**0.10**

the supplementary material. We empirically evaluate the suitability of hashtags in representing the topics and found that, consistently, over 90% of the hashtags are good candidates for the topic labels.

Inference on Authors' Topic Distributions In addition to inference on the topic distribution of each document, the TN topic model allows us to analyze the topic distribution of each author. Table 3 presents a summary of topics by different authors, where topics are obvious from the Twitter ID.

Author Recommendation We illustrate the use of the TN topic model for author recommendation. On a new test dataset with 90451 tweets and 625 new authors, we predict the most similar and dissimilar authors for the new authors, based on the training model of 60370 tweets. We quantify the recommendation quality with the cosine similarities of the authors' topic distributions for the recommended author pairs. We compare our new kernel function with the original kernel function (denoted as *original*) used in [16]. Table 4 shows average cosine similarities between the recommended and not-recommended authors. This suggests that our kernel function is more appropriate. Additionally, we manually checked the recommended authors and we found that they usually belong to the same community, *i.e.*, having tweets with similar topics.

Clustering and Topic Coherence We also evaluate the TN topic model against state-of-the-art LDA-based clustering techniques [21]. We find that the TN topic model outperforms the state-of-the-art in *purity*, normalized mutual information and pointwise mutual information (PMI). Due to space, the evaluation result is provided in the supplementary material.

6 Conclusion and Future Work

We propose a full Bayesian nonparametric *Twitter-Network* (TN) topic model that jointly models tweets and the associated social network information. Our model employs a nonparametric Bayesian approach by using the PDP and GP, and achieves flexible modeling by performing inference on a network of PDPs. Our experiments with Twitter dataset show that the TN topic model achieves significant improvement compared to existing baselines. Furthermore, our ablation study demonstrates the usefulness of each component of the TN model. Our model also shows interesting applications such as *author recommendation*, as well as providing additional informative inferences.

We also engineered a framework for rapid topic model development, which is important due to the complexity of the model. While we could have used Adaptor Grammars [22], our framework yields more efficient computation for topic models.

Future work includes speeding up the posterior inference algorithm, especially for the network model, as well as incorporating other auxiliary information that is available in social media such as *location*, *hyperlinks* and *multimedia contents*. We also intend to explore other applications that can be addressed with the TN topic model, such as *hashtag recommendation*. It is also interesting to apply the TN topic model to other types of data such as blog and publication data.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful feedback and comments.

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

[2] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. ECIR, 2011.

[3] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. How noisy social media text, how diffrnt social media sources? In *IJCNLP*, 2013.

[4] J. Eisenstein. What to do about bad language on the internet. In *NAACL*. ACL, 2013.

[5] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[6] F. S. Tsai. A tag-topic model for blog mining. *Expert Syst. Appl.*, 38(5):5330–5335, 2011.

[7] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link LDA: joint models of topic and author community. In *ICML*, 2009.

[8] M. Qiu, F. Zhu, and J. Jiang. It is not just what we say, but how we say them: LDA-based behavior-topic model. In *SDM*, 2013.

[9] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. WWW '13, 2013.

[10] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.

[11] S. Goldwater, T. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. *NIPS*, 18:459, 2006.

[12] I. Sato and H. Nakagawa. Topic models with power-law using Pitman-Yor process. KDD '10, pages 673–682. ACM, 2010.

[13] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Ann. Appl. Stat.*, 4(1):124–150, 2010.

[14] R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *KDD*, 2008.

[15] K. Miller, M. Jordan, and T. Griffiths. Nonparametric latent feature models for link prediction. In *NIPS*, pages 1276–1284, 2009.

[16] J. Lloyd, P. Orbanz, Z. Ghahramani, and D. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *NIPS*, 2012.

[17] W. Buntine, L. Du, and P. Nurmi. Bayesian networks on Dirichlet distributed vectors. *Proceedings of the fifth European workshop on probabilistic graphical models*, 2010.

[18] C. Chen, L. Du, and W. Buntine. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *ECML*. 2011.

[19] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, 2011.

[20] M. Lui and T. Baldwin. langid.py: an off-the-shelf language identification tool. In *ACL*, pages 25–30, 2012.

[21] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*, 2013.

[22] M. Johnson, T. L. Griffiths, and S. Goldwater. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. *NIPS*, 19:641, 2007.