Markov Random Topic Fields

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, UT 84112
me@hal3.name

Abstract

Most approaches to topic modeling assume an independence between documents that is frequently violated. We present an topic model that makes use of one or more user-specified graphs describing relationships between documents. These graph are encoded in the form of a Markov random field over topics and serve to encourage related documents to have similar topic structures. Experiments on show upwards of a 10% improvement in modeling performance.

1 Introduction

One often wishes to apply topic models to large document collections. In these large collections, we usually have meta-information about how one document relates to another. Perhaps two documents share an author; perhaps one document cites another; perhaps two documents are published in the same journal or conference. We often believe that documents related in such a way should have similar topical structures. We encode this in a probabilistic fashion by imposing an (undirected) Markov random field (MRF) on top of a standard topic model (see Section 3). The edge potentials in the MRF encode the fact that "connected" documents should share similar topic structures, measured by some parameterized distance function. Inference in the resulting model is complicated by the addition of edge potentials in the MRF. We demonstrate that a hybrid Gibbs/Metropolis-Hastings sampler is able to efficiently explore the posterior distribution (see Section 4).

In experiments (Section 5), we explore several variations on our basic model. The first is to explore the importance of being able to tune the strength of the potentials in the MRF as part of the inference procedure. This turns out to be of utmost importance. The second is to study the importance of the form of the distance metric used to specify the edge potentials. Again, this has a significant impact on performance. Finally, we consider the use of multiple graphs for a single model and find that the power of combined graphs also leads to significantly better models.

2 Background

Probabilistic topic models propose that text can be considered as a mixture of words drawn from one or more "topics" (Deerwester et al., 1990; Blei et al., 2003). The model we build on is latent Dirichlet allocation (Blei et al., 2003) (henceforth, LDA). LDA stipulates the following generative model for a document collection:

1. For each document $d = 1 \dots D$:
   (a) Choose a topic mixture $\boldsymbol{\theta}_d \sim \mathcal{D}ir(\boldsymbol{\alpha})$
   (b) For each word in d, $n = 1 \dots N_d$:
      i. Choose a topic $z_{dn} \sim \mathcal{M}ult(\boldsymbol{\theta}_d)$
      ii. Choose a word $w_{dn} \sim \mathcal{M}ult(\boldsymbol{\beta}_{z_{dn}})$

Here, $\boldsymbol{\alpha}$ is a hyperparameter vector of length K, where K is the desired number of topics. Each document has a topic distribution $\boldsymbol{\theta}_d$ over these K topics and each word is associated with precisely one topic (indicated by z_{dn}). Each topic $k = 1 \dots K$ is a unigram distribution over words (aka, a multinomial) parameterized by a vector $\boldsymbol{\beta}_k$. The associated graphical model for LDA is shown in Figure 1. Here, we have added a few additional hyperparameters: we place a $\mathcal{G}am(a, b)$ prior independently on each component of $\boldsymbol{\alpha}$ and a $\mathcal{D}ir(\eta, \dots, \eta)$ prior on each of the $\boldsymbol{\beta}$s.

The joint distribution over all random variables specified by LDA is:

$$p(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{w}) = \prod_k \mathcal{G}am(\alpha_k \mid a, b)\mathcal{D}ir(\boldsymbol{\beta}_k \mid \eta) \qquad (1)$$
$$\prod_d \mathcal{D}ir(\theta_d \mid \boldsymbol{\alpha}) \prod_n \mathcal{M}ult(z_{dn} \mid \theta_d)\mathcal{M}ult(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}})$$

Many inference methods have been developed for this model; the approach upon which we

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 293–296,
Suntec, Singapore, 4 August 2009.

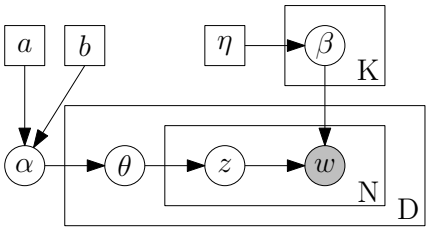

Figure 1: Graphical model for LDA.

build is the collapsed Gibbs sampler (Griffiths and Steyvers, 2006). Here, the random variables $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are analytically integrated out. The main sampling variables are the z_{dn} indicators (as well as the hyperparameters: η and a, b). The conditional distribution for z_{dn} conditioned on all other variables in the model gives the following Gibbs sampling distribution $p(z_{dn} = k)$:

$$\frac{\#_{z=k}^{-dn} + \alpha_k}{\sum_{k'}(\#_{z=k'}^{-dn} + \alpha_{k'})} \frac{\#_{z=k,w=w_{dn}}^{-dn} + \eta}{\sum_{k'}(\#_{z=k',w=w_{dn}}^{-dn} + \eta)} \quad (2)$$

Here, $\#_{\chi}^{-dn}$ denotes the number of times event χ occurs in the entire corpus, excluding word n in document d. Intuitively, the first term is a (smoothed) relative frequency of topic k occurring; the second term is a (smoothed) relative frequency of topic k giving rise to word w_{dn}.

A Markov random field specifies a joint distribution over a collection of random variables $x_1, \ldots, x_N$. An undirected graph structure stipulates how the joint distribution factorizes over these variables. Given a graph $\mathcal{G} = (V, E)$, where $V = \{x_1, \ldots, x_N\}$, let $\mathcal{C}$ denote a subset of all the cliques of $\mathcal{G}$. Then, the MRF specifies the joint distribution as: $p(\boldsymbol{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\boldsymbol{x}_c)$. Here, $Z = \sum_{\boldsymbol{x}} \prod_{c \in \mathcal{C}} \psi_c(\boldsymbol{x}_c)$ is the partition function, $\boldsymbol{x}_c$ is the subset of $\boldsymbol{x}$ contained in clique c and ψ_c is *any* non-negative function that measures how "good" a particular configuration of variables $\boldsymbol{x}_c$ is. The ψs are called potential functions.

3 Markov Random Topic Fields

Suppose that we have access to a collection of documents, but do not believe that these documents are all independent. In this case, the generative story of LDA no longer makes sense: related documents are more likely to have "similar" topic structures. For instance, in the scientific community, if paper A cites paper B, we would (a priori) expect the topic distributions for papers A and B to be related. Similarly, if two papers share an author, we might expect them to be topically related.

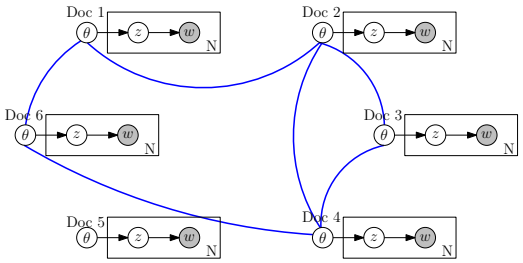

Figure 2: Example Markov Random Topic Field (variables α and $\boldsymbol{\beta}$ are excluded for clarify).

Of if they are both published at EMNLP. Or if they are published in the same year, or come out of the same institution, or many other possibilities.

Regardless of the source of this notion of similarity, we suppose that we can represent the relationship between documents in the form of a graph $\mathcal{G} = (V, E)$. The vertices in this graph are the documents and the edges indicate relatedness. Note that the resulting model will not be fully generative, but is still probabilistically well defined.

3.1 Single Graph

There are multiple possibilities for augmenting LDA with such graph structure. We could "link" the topic distributions $\boldsymbol{\theta}$ over related documents; we could "like" the topic indicators $\boldsymbol{z}$ over related documents. We consider the former because it leads to a more natural model. The idea is to "unroll" the D-plate in the graphical model for LDA (Figure 1) and connect (via undirected links) the $\boldsymbol{\theta}$ variables associated with connected documents. Figure 2 shows an example MRTF over six documents, with thick edges connecting the $\boldsymbol{\theta}$ variables of "related" documents. Note that each $\boldsymbol{\theta}$ still has α as a parent and each w has $\boldsymbol{\beta}$ as a parent: these are left off for figure clarity.

The model is a straightforward "integration" of LDA and an MRF specified by the document relationships $\mathcal{G}$. We begin with the joint distribution specified by LDA (see Eq (1)) and add in edge potentials for each edge in the document graph $\mathcal{G}$ that "encourage" the topic distributions of neighboring documents to be similar. The potentials all have the form:

$$\psi_{d,d'}(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) = \exp\left[-\ell_{d,d'} \rho(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'})\right] \quad (3)$$

Here, $\ell_{d,d'}$ is a "measure of strength" of the importance of the connection between d and d' (and will be inferred as part of the model). ρ is a distance metric measuring the dissimilarity between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_{d'}$. For now, this is Euclidean distance

(i.e., $\rho(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) = ||\boldsymbol{\theta}_d - \boldsymbol{\theta}_{d'}||$); later, we show that alternative distance metrics are preferable.

Adding the graph structure necessitates the addition of hyperparameters ℓ_e for every edge $e \in E$. We place an exponential prior on each $1/\ell_e$ with parameter λ: $p(\ell_e \mid \lambda) = \lambda \exp(-\lambda/\ell_e)$. Finally, we place a vague $\mathcal{G}am(\lambda_a, \lambda_b)$ prior on λ.

3.2 Multiple Graphs

In many applications, there may be multiple graphs that apply to the same data set, $\mathcal{G}_1, \ldots, \mathcal{G}_J$. In this case, we construct a single MRF based on the union of these graph structures. Each edge now has L-many parameters (one for each graph j) ℓ^j_e. Each graph also has its own exponential prior parameter λ_j. Together, this yields:

$$\psi_{d,d'}(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'}) = \exp\left[-\sum_j \ell^j_{d,d'} \rho(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'})\right] \quad (4)$$

Here, the sum ranges only over those graphs that have (d, d') in their edge set.

4 Inference

Inference in MRTFs is somewhat complicated from inference in LDA, due to the introduction of the additional potential functions. In particular, while it is possible to analytically integrate out $\boldsymbol{\theta}$ in LDA (due to multinomial/Dirichlet conjugacy), this is no longer possible in MRTFs. This means that we must explicitly represent (and sample over) the topic distributions $\boldsymbol{\theta}$ in the MRTF.

This means that we must sample over the following set of variables: $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, $\boldsymbol{z}$, $\boldsymbol{\ell}$ and λ. Sampling for α remains unchanged from the LDA case. Sampling for variables *except* $\boldsymbol{\theta}$ is easy:

$$z_{dn} = k: \quad \theta_{dk} \frac{\#^{-dn}_{z=k, w=w_{dn}} + \eta}{\sum_{k'} (\#^{-dn}_{z=k', w=w_{dn}} + \eta)} \quad (5)$$

$$1/\ell_{d,d'} \sim \mathcal{E}xp\Big(\lambda + \rho(\boldsymbol{\theta}_d, \boldsymbol{\theta}_{d'})\Big) \quad (6)$$

$$\lambda \sim \mathcal{G}am\Big(\lambda_a + |E|, \lambda_b + \sum_e \ell_e\Big) \quad (7)$$

The latter two follow from simple conjugacy. When we use multiple graphs, we assign a separate λ for each graph.

For sampling $\boldsymbol{\theta}$, we resort to a Metropolis-Hastings step. Our proposal distribution is the Dirichlet *posterior* over $\boldsymbol{\theta}$, given all the current assignments. The acceptance probability then just depends on the graph distances. In particular, once $\boldsymbol{\theta}_d$ is drawn from the posterior Dirichlet, the acceptance probability becomes $\prod_{d' \in \mathcal{N}(d)} \psi_{d,d'}$, where $\mathcal{N}(d)$ denotes the neighbors of d. For each document, we run 10 Metropolis steps; the acceptance rates are roughly 25%.

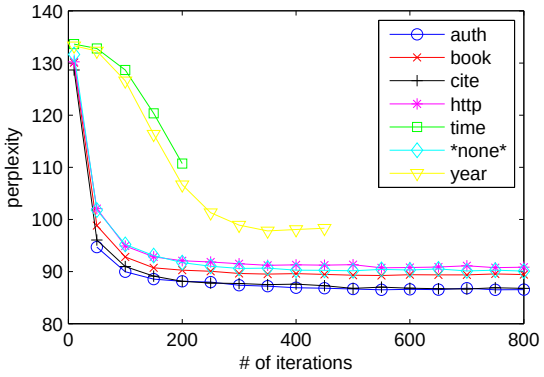

Figure 3: Held-out perplexity for different graphs.

5 Experiments

Our experiments are on a collection for 7441 document abstracts crawled from CiteSeer. The crawl was seeded with a collection of ten documents from each of: ACL, EMNLP, SIGIR, ICML, NIPS, UAI. This yields 650 thousand words of text after remove stop words. We use the following graphs (number in parens is the number of edges):

auth: shared author ($47k$)
book: shared booktitle/journal ($227k$)
cite: one cites the other ($18k$)
http: source file from same domain ($147k$)
time: published within one year ($4122k$)
year: published in the same year ($2101k$)

Other graph structures are of course possible, but these were the most straightforward to cull.

The first thing we look at is convergence of the samplers for the different graphs. See Figure 3. Here, we can see that the author graph and the citation graph provide improved perplexity to the straightforward LDA model (called "*none*"), and that convergence occurs in a few hundred iterations. Due to their size, the final two graphs led to significantly slower inference than the first four, so results with those graphs are incomplete.

Tuning Graph Parameters. The next item we investigate is whether it is important to tune the graph connectivity weights (the ℓ and λ variables). It turns out this is *incredibly* important; see Figure 4. This is the same set of results as Figure 3, but without ℓ and λ tuning. We see that the graph-based methods *do not* improve over the baseline.

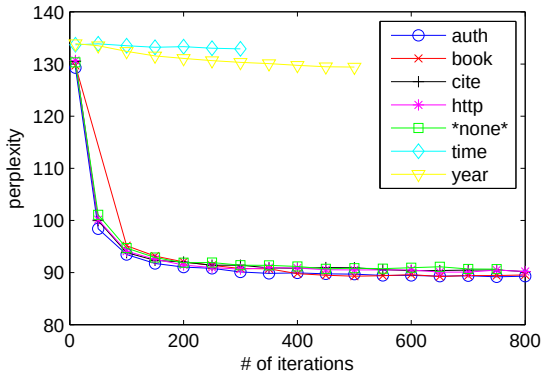

Figure 4: Held-out perplexity for difference graph structures without graph parameter tuning.

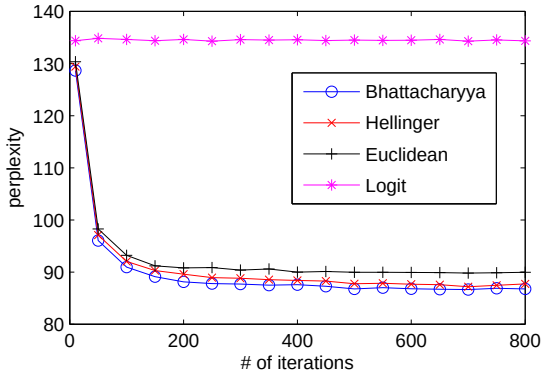

Figure 5: Held-out perplexity for different distance metrics.

Distance Metric. Next, we investigate the use of different distance metrics. We experiments with Bhattacharyya, Hellinger, Euclidean and logistic-Euclidean. See Figure 5 (this is just for the auth graph). Here, we see that Bhattacharyya and Hellinger (well motivated distances for probability distributions) outperform the Euclidean metrics.

Using Multiple Graphs Finally, we compare results using combinations of graphs. Here, we run every sampler for 500 iterations and compute standard deviations based on ten runs (year and time are excluded). The results are in Table 1. Here, we can see that adding graphs (almost) always helps and never hurts. By adding all the graphs together, we are able to achieve an absolute reduction in perplexity of 9 points (roughly 10%). As discussed, this hinges on the tuning of the graph parameters to allow different graphs to have different amounts of influence.

6 Discussion

We have presented a graph-augmented model for topic models and shown that a simple combined Gibbs/MH sampler is efficient in these models.

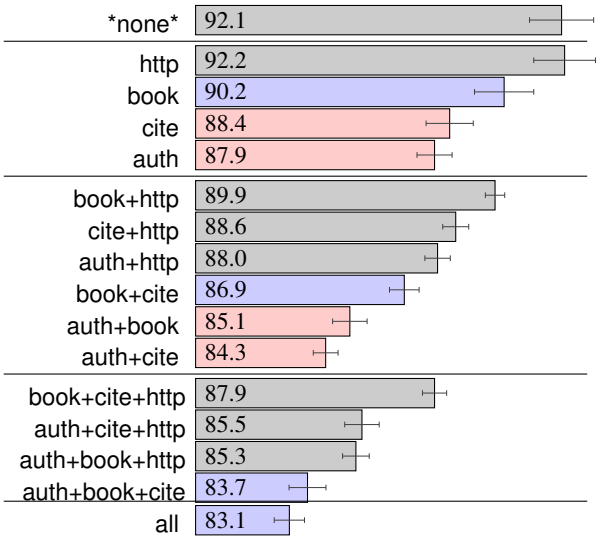

Table 1: Comparison of held-out perplexities for varying graph structures with two standard deviation error bars; grouped by number of graphs. Grey bars are indistinguishable from best model in previous group; blue bars are at least two stddevs better; red bars are at least four stddevs better.

Using data from the scientific domain, we have shown that we can achieve significant reductions in perplexity on held-out data using these models. Our model resembles recent work on hypertext topic models (Gruber et al., 2008; Sun et al., 2008) and blog influence (Nallapati and Cohen, 2008), but is specifically tailored toward undirected models. Ours is an alternative to the recently proposed Markov Topic Models approach (Wang et al., 2009). While the goal of these two models is similar, the approaches differ fairly dramatically: we use the graph structure to inform the per-document topic distributions; they use the graph structure to inform the unigram models associated with each topic. It would be worthwhile to directly compare these two approaches.

References

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *JMLR*, 3.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6).

Tom Griffiths and Mark Steyvers. 2006. Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*.

Amit Gruber, Michal Rosen-Zvi, , and Yair Weiss. 2008. Latent topic models for hypertext. In *UAI*.

Ramesh Nallapati and William Cohen. 2008. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Conference for Webblogs and Social Media*.

Congkai Sun, Bin Gao, Zhenfu Cao, and Hang Li. 2008. HTM: A topic model for hypertexts. In *EMNLP*.

Chong Wang, Bo Thiesson, Christopher Meek, and David Blei. 2009. Markov topic models. In *AI-Stats*.