A simple non-parametric Topic Mixture for Authors and Documents

Arnim Bleier

GESIS - Leibniz Institute for the Social Sciences
Knowledge Technologies for Social Sciences
Unter Sachsenhausen 6-8, 50667 Cologne, Germany
arnim.bleier@gesis.org

Abstract. This article reviews the Author-Topic Model and presents a new non-parametric extension based on the Hierarchical Dirichlet Process. The extension is especially suitable when no prior information about the number of components necessary is available. A blocked Gibbs sampler is described and focus put on staying as close as possible to the original model with only the minimum of theoretical and implementation overhead necessary.

Keywords: Topic models, Author models, Dirichlet process, Stick-breaking Prior, Markov chain Monte Carlo

1 Introduction

Probabilistic models to infer the interests of authors have attracted much interest throughout the language modeling community, with the Author-Topic model [4] as one of its seminal representatives. Multiple modifications to the Author-Topic model have been proposed. These modifications assume either a fixed number of topics or focus on using authorship information as an additional feature in a non-parametric setting with only little resemblance to the structure of the original work. This article addresses a complementary problem – representing the Author-Topic model in the framework of Bayesian non-parametrics but keeping as much as possible of its original structure. While this might be valuable in its own right, it is also useful in a more general sense since the steps necessary to transform an extension of Latent Dirichlet Allocation (LDA) with a fixed number of parameters to an equivalent model that grows the number of parameters with the amount data available apply to a broad range of models.

2 Generative models for documents and authors

We will describe two different models: The first one relates authors and documents via a fixed number of topics, and the second one models the interests of authors using a flexible number of topics. Both models are described by using the common notation of a document d being a vector of N_d words, $\mathbf{w_d}$, where

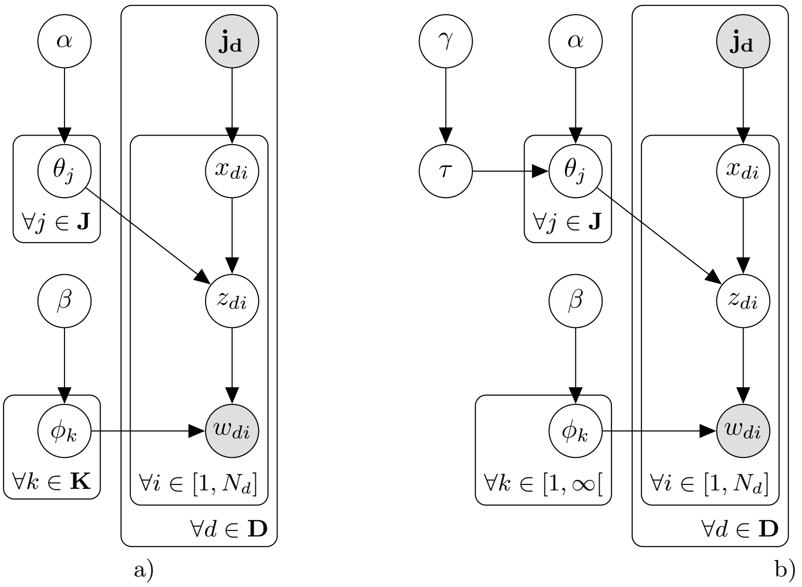

Fig. 1. Admixture models for documents and authors: (a) The Author-Topic model, (b) the non-parametric Author-Topic model (this paper).

all w_{di} are chosen from a vocabulary with V terms, and $\mathbf{j_d}$ are the authors of document d chosen from the set of all authors of size J. A corpus of D documents is then defined by the set $\{(\mathbf{w_1}, \mathbf{j_1}), .., (\mathbf{w_D}, \mathbf{j_D})\}$.

2.1 The parametric model

The seminal Author-Topic model [4] has two sets of unknown parameters; J distributions θ_j over topics conditioned on the authors, and K distributions ϕ_k over terms conditioned on the topics - as well as the assignments of individual words to authors x_{di} and topics z_{di}. With θ and ϕ being integrated out a collapsed gibbs sampler is used, analogous to [1], to converge to the true underlying distributions of the Markov state variables x and z. The transitions between the states of the chain result from iteratively sampling each pair (x_{di}, z_{di}) as a block, conditioned on all other variables:

$$p(z_{di} = k, x_{di} = j \mid w_{di} = t, \mathbf{w}_{-di}, \mathbf{j}_d, \cdot) \propto \frac{n_{jk}^{-di} + \alpha}{n_{j.}^{-di} + K\alpha} f_{k_{di}}^{-w_{di}}(t) \qquad (1)$$

where n_{jk}^{-di} is the number of times a word of the topic k has been assigned to the author j excluding the current instance, and $\cdot$ is used in place of a variable to indicate that the sum over its values (e.g. $n_{j.} = \sum_k n_{jk}$) is taken. The assignment variable $z_{di} = k$ represents the topic of the i^{th} word in document d being k as

$x_{di} = j$ represents the assignment to author j. The term on the right side $f_{k_{di}}^{-w_{di}}(t)$ is the posterior density of term t under topic k:

$$f_{k_{di}}^{-w_{di}}(t) = \frac{n_{kt}^{-di} + \beta}{n_{k.}^{-di} + V\beta} \tag{2}$$

where n_{kt}^{-di} is the number of times a term t has been assigned to topic k again excluding the current word from the count.

2.2 The non-parametric model

One frequently raised question when applying the Author-Topic model to a new data set, is how to choose the number of topics [5]. The Bayesian non-parametric framework of the Hierarchical Dirichlet Process (HDP) [6] offers an elegant solution to this by allowing a prior over a countably infinite number of topics of which only a few will dominate the posterior. Building on the finite version of the model we split the symmetric prior α over topics into a scalar precision parameter α and a distribution $\tau \sim Dir(\gamma/K)$. Taking this to the limit $K \to \infty$ we get the root distribution for the non-parametric Author-Topic model (fig.1b). Analogously to the collapsed gibbs sampler for the previous LDA version we integrate over θ_j, but keep τ as an auxiliary variable to preserve the structure of the state transition probabilities in the finite case for the HDP [2].

$$p(z_{di} = k, x_{di} = j \mid w_{di} = t, \tau, \mathbf{j}_d, \cdot) \propto \begin{cases} \dfrac{n_{jk} + \alpha\tau_k}{n_{j.} + \alpha} f_{k_{di}}^{-w_{di}}(t) & \text{if } z = k \\ \dfrac{\alpha\tau_{k^{+1}}}{n_{j.} + \alpha} f_{k^{new}}^{-w_{di}}(t) & \text{if } z = k_{new} \end{cases} \tag{3}$$

With $f_{k^{new}}^{-w_{di}}(t) = \frac{1}{V}$ being the prior density of a word w under a new topic [6]. The key difference between these equations and the original model (1) is that we now have a root distribution τ for the HDP over K+1 possible states. If there are K topics in the current step, then $\tau_{k^{+1}}$ represents the accumulated continuous probability mass of all possible but currently unused topics, allowing to choose a new one from a countably infinite pool of empty topics. If the count for number of words assigned to a topic goes to zero, the topic is returned to the pool of unused topics.

2.3 Sampling the Root Distribution

However, the construction of a Markov chain for the non-parametric Author-Topic model requires that additionally the root distribution τ of the Dirichlet processes must be sampled which was not present in the finite version of the model. The discrete part of the root distribution guarantees that existing topics are reused with probability $\sum_K \tau_k$ and the continuous part allows for a new topic to be sampled with probability τ_{k+1} [3]. Given the Markov state we begin by generating J vectors

$$\mu_{jkr} = \frac{\alpha\tau_k}{r - 1 + \alpha\tau_k}; \text{ with } r = 1, .., n_{jk} \tag{4}$$

where n_{jk} are the number of words for author j which have been assigned to topic k. Next, we draw Bernoulli random variables $m_{jkr} \sim Bern(\mu_{jkr})$. The posterior of the top-level Dirichlet process τ is then sampled via

$$\tau \sim Dir([m_1, .., m_k], \gamma); \text{ with } m_k = \sum_{jr} m_{jkr} \tag{5}$$

making τ a discrete distribution over K used topics plus one component with the probability mass of the infinite possible, yet unused topics.

3 Discussion

In this work, we transformed the LDA based Author-Topic model into a non-parametric model that estimates the number of components necessary for representing the data. Yet, it will be necessary to empirically evaluate performance (i.e. perplexity) of the proposed model on benchmark data sets. While choosing the Author-Topic model as an example for such a transformation, we believe that many of the considerations made equally hold for a wider range of models and can serve as a blueprint for a simple application of non-parametric Bayesian priors.

References

[1] Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(1), 5228–5235 (2004)

[2] Heinrich, G.: “Infinite LDA” – Implementing the HDP with minimum code complexity. Technical report, arbylon.net (2011)

[3] Porteous, I.: Mixture Block Methods for Non Parametric Bayesian Models with Applications. Ph.D. thesis (2010)

[4] Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The Author-Topic Model for Authors and Documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (2004)

[5] Sugimoto, C., Li, D., Russell, T., Finlay, S., Ding, Y.: The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. Journal of the American Society for Information Science and Technology 62(1), 185–204 (2011)

[6] Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet Processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)