Modeling corpora of timestamped documents using semisupervised nonparametric topic models

Ramnath Balasubramanyan
Language Technologies Institute
Carnegie Mellon University
rbalasub@cs.cmu.edu

William W. Cohen
Machine Learning Department
Carnegie Mellon University
wcohen@cs.cmu.edu

Matthew Hurst
Microsoft Corporation
mhurst@microsoft.com

Abstract

In this paper we propose a nonparametric topic model to capture the evolution of text over time. Mixture models for modeling text documents based on hierarchical Dirichlet processes (HDP) have been used successfully in recent work to provide a nonparametric prior for the number of topics in the corpus eliminating the need to specify *apriori* the number of topics. We extend this model to additionally model timestamps associated with documents using Gaussian distributions to make the induced topics time-sensitive, thus modeling dynamic structure in the corpus. We present the new model, hierarchical Dirichlet process over time (HOT), in the framework of a Chinese restaurant franchise process, and describe a Markov chain Monte Carlo algorithm for performing approximate posterior inference. We demonstrate the capability of the HOT model to capture temporally varying structure in two corpora of blog posts, and a third corpus of NIPS abstracts. Experiments show that our new model performs as well if not better than its best hand-tuned parametric counterparts with respect to two measures: document perplexity, and prediction of the timestamp of documents from its content. We also describe a framework to provide human guidance to the topic induction process and to adapt the MCMC sampling technique into a semisupervised procedure. Experiments show that even limited human guidance improves the effectiveness of the model.

1 Introduction

Topic models such as latent Dirichlet allocation (LDA) [1] have become a common way to describe documents in a low-dimensional space. When used to model text documents, LDA represents each document in a corpus as a finite mixture over underlying "topics", and each topic is in turn represented as a distribution over terms in the dictionary. Efficient sampling and variational inference algorithms [2] are used to perform approximate inference on the posterior distributions of the topic probabilities of documents and the underlying set of topics.

Time-dependent extensions of LDA such as topics over time (TOT)[3] and multi-scale topic tomography[4] also model temporal structures in a corpus that change over time. TOT parameterizes topics with a continuous Beta distribution over "timestamps" in a document in addition to a distribution over words. Topics that are frequent in a corpus over a narrow time span are represented by Beta distributions which are peaked, and persistent topics are represented by flatter time distributions. Explicitly modeling time helps in separating distinct topics that share common words

but occur at different time periods: for instance, TOT is well suited to separate out the Iraq wars of 1991 and 2003 as separate topics in a news corpus that spans from 1989 to 2005, which would not be possible in LDA barring the presence of a strong signal in the vocabulary which distinguishes the two wars.

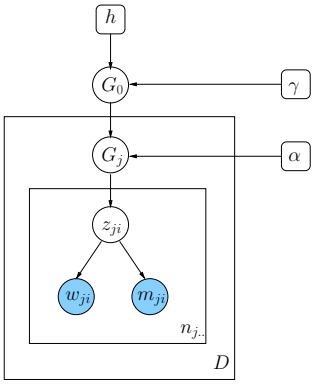
When applying TOT to certain corpora (such as the collections of political blog posts considered below), peaked topics often correspond to widely-discussed external events. In general, the number of such events is difficult to predict *a priori*; more generally, often the appropriate number of topics to use in a model is unknown. One solution to this problem is to use cross validation to pick the best number of clusters using evaluation metrics like document perplexity. An elegant alternative is the use of nonparametric priors such as Dirichlet processes [5, 6]. With Dirichlet processes, the exact number of topics created is dependent on the corpus and the value of the hyperparameters of the process. It can be argued that this approach is more efficient, and more systematic, than using cross validation to find the optimal number of topics, since the search for the number of topics is interleaved with the search for topic parameters, whereas use of cross-validation requires repeatedly running an LDA-like model to convergence, with a different number of topics.

Hierarchical Dirichlet processes (HDP) [7] use a hierarchy of Dirichlet processes to model groups of observed data that are linked by a common structure. Further work by the same authors describes how they can be used as priors in a mixture model to create an "infinite" topic model [8]. In this paper, we present hierarchical Dirchlet process over time (HOT), a hierarchical mixture model to capture evolving structure in a set of documents. A hierarchical Dirichlet process is used as a prior over the number of topics. Most importantly, the topics themselves explicitly generate timestamps of documents using Gaussian distributions in addition to modeling terms in documents. In the generative process described by the model, a top level Dirichlet process provides a prior distribution for the number of underlying topics in the corpus. Draws from this process supply an underlying countably infinite set of topics. Each document in turn has a Dirichlet process prior which dictates the number of topics that its topic distribution spans. Since the support of the process is the set of draws from the top level Dirichlet process, the documents in the corpus share topics amongst themselves. Each word in a document is created by drawing a topic from its topic distribution and subsequently generating a word from the vocabulary along with a timestamp for the document. A timestamp is drawn once per word in the document to keep the sampling algorithm simple. Since timestamps are observed variables, we enforce the constraint that the document timestamp is used as the one generated for every word.

One intuitive way to look at Dirichlet processes is the Chinese restuarant process [9]. Teh et al. describe hierarchical Dirichlet processes using a Chinese restaurant franchise to handle the extra layer of complexity introduced by hierarchical stacking of Dirichlet processes. In this paper we use the Chinese restaurant franchise representation of HDP to devise a simple MCMC sampling technique to perform posterior inference on the topic probabilities and the the distributions over timestamps and words in topics.

Topic models have the desirable properties of being wholly unsupervised. While this is advantageous by obviating the need for expensive human annotation, a system for introducing human input can be useful as evidenced by work such as [10, 11]. We describe a method for introducing human input into HOT by partially describing topics using a few keywords and/or an approximate timestamp associated with it. The partial description of a topic provided as input is assumed to be a part a *pseudo-document* associated with it. The total probability mass of the topic probability distribution of the pseudo document resides in the topic which it represents. During inference, topic membership for words in the document are locked and not sampled by the MCMC algorithm. This framework enables human input without major changes in the inference algorithm.

The paper is organized as follows. In section 2 we describe the parameterization of the HOT model, the MCMC sampling algorithm to perform posterior inference and the framework to provide user input. Section 3 gives details on the three datasets used in the experimental section. Results from the experiments are presented and discussed in section 4. We finally present our conclusions in section 5.



h - $\langle \text{DirSym}(\beta), \text{InverseGamma}(\upsilon, \zeta^2) \rangle$; topic priors

$G_0 \sim DP(\gamma, h)$

$G_j \sim DP(\alpha, G_0)$

$z_{ji} \sim G_j$ - topic drawn for generating word w_{ji} and timestamp m_{ji} in ith word of the jth document.

$\phi_1, \phi_2, \ldots, \phi_K$ - draws instantiated from G_0 (The countably infinite set of topics available to documents)

$\psi_{j1}, \psi_{j2}, \ldots, \psi_{jK_j}$ - draws instantiated from G_j (The subset of topics that are used in document j)

D - number of documents in the corpus.

Figure 1: Hierarchical Dirichlet process mixture model over time

2 Model

We propose a nonparametric topic model that generates documents with timestamps. This model is based on the hierarchical Dirichlet process mixture model [8]. The difference in this model from HDP lies in the additional components added to generate the timestamps of documents.

2.1 Parameters

A topic ϕ in the model is represented by a pair $\langle \Theta, \xi \rangle$. Θ is a multinomial which determines the words generated by the topic and is parameterized by $(\theta_1, \theta_2, \ldots, \theta_{|V|})$ where V is the vocabulary. ξ is a Gaussian $N(\mu, \sigma^2)$ which emits topic specific timestamps.

It can be seen in Fig 1 that the timestamp for the document is generated once per word instead of once per document. This is to simplify the MCMC sampling equations and does not cause problems since the timestamps are observed and constrained to be the same for all the words. The impact of timestamps can be controlled by weighting the observed variables as desired.

Each level of the hierarchical Dirichlet process provides a nonparametric prior over the space of multinomials and Gaussians.

2.2 Approximate inference using Monte Carlo Markov Chain (MCMC) sampling

Exact inference for the model described above is intractable. Therefore, we use a MCMC-based method to perform approximate inference. The sampling method is adapted from the Chinese Restaurant Process [7]. Every word in the document is deemed to be generated from a table which in turn is assigned a dish from a global set of dishes. Having a global set of dishes which represent topics ensures that topics are shared between documents.

In order to make the sampling straighforward, we use a set of index variables instead of directly sampling z_{ji} or ψ. The index variables used are

- t_{ji} - index into $\psi_{j1}, \psi_{j2}, \ldots, \psi_{jk_j}$, table number assigned to word i in the jth document i.e $z_{ji} = \psi_{jt_{ji}}$.
- k_{jt} - dish number assigned to the topic t in document j i.e. $\psi_{jt} = \phi_{k_{jt}}$.
- Finally, for convenience we collapse t and k into $u_{ji} = k_{jt_{ji}}$

We also need variables to represent counts.

- n_{jtk} - number of words in document j assigned to table t and dish k.

- m_{jk} - number of tables in document j that serve dish k.

Dots in the subscripts of the count variables represent marginal counts, marginalized over the variable replaced with a dot. When a superscript is attached to a set of variables or a count, the variable corresponding to the superscripted index is removed from the set used in the calculation of the count.

We next define some quantities that are useful when deriving the sampling equations.

$$
\begin{aligned}
c_{k,w} &= \sum_{\substack{j'i' \neq ji, \\ u_{j'i'}=k}} I(w_{j'i'} = w) \\
f_k^{ji} &= p(w_{ji}, m_{ji} | u_{ji} = k, \mathbf{w}^{-\mathbf{ji}}, \mathbf{m}^{-\mathbf{ji}}, \mathbf{u}^{-\mathbf{ji}}) \\
&= \frac{\int p(w_{ji}, m_{ji}|\phi_k) \prod_{\substack{j'i' \neq ji, \\ u_{j'i'}=k}} p(w_{j'i'}, m_{j'i'}|\phi_k) h(\phi_k)\, d\phi_k}{\int \prod_{\substack{j'i' \neq ji, \\ u_{j'i'}=k}} p(w_{j'i'}, m_{j'i'}|\phi_k) h(\phi_k)\, d\phi_k} \\
&= \frac{c_{k,w_{ji}} + \beta}{\sum_{w \in V}(c_{k,w} + \beta)} \times \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(\frac{-(m_{ji} - \hat{\mu})^2}{2\hat{\sigma}^2} \right)
\end{aligned}
$$

where $\hat{\mu} = \frac{\sum_{\substack{j'i' \neq ji, \\ u_{ji}=k}} m_{j'i'}}{n_{..k}}$, and $\hat{\sigma}^2 = \frac{2\zeta^2 + \sum_{\substack{j'i' \neq ji, \\ u_{ji}=k}} (m_{j'i'} - \hat{\mu})^2}{(2(v+1)+n_{..k})}$.

For a new dish, i.e. when $k = k^{new}$, $f_{k^{new}}^{ji} = \int p(w_{ji}|\Theta) p(m_{ji}|\xi) h(\phi)\, d\phi = \frac{1}{|V|}$ reduces to the prior density of $\langle w_{ji}, m_{ji} \rangle$.

2.2.1 Sampling t

The table assignments to each term in the document is governed by the distribution.

$$
p(t_{ji} = t | \mathbf{t}^{-\mathbf{ji}}, \mathbf{k}) \propto \begin{cases} n_{jt.}^{-ji} f_{k_{jt}}^{ji} & \text{if } t \text{ previously used,} \\ \alpha\, p(w_{ji}, m_{ji} | \mathbf{t}^{-\mathbf{ji}}, t_{ji} = t^{new}, \mathbf{k}) & \text{if } t = t^{new} \end{cases}
$$

Here $p(w_{ji}, m_{ji} | \mathbf{t}^{-\mathbf{ji}}, t_{ji} = t^{new}, \mathbf{k})$ is the probability of the word when a new table is opened is given by $\sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\gamma} f_k^{ji} + \frac{\gamma}{m_{..}+\gamma} f_{k^{new}}^{ji}$

If sampled value of t_{ji} is t^{new}, we obtain a sample of $k_{jt^{new}}$ by sampling

$$
p(k_{jt^{new}} = k | \mathbf{t}, \mathbf{k}^{-jt^{new}}) \propto \begin{cases} m_{.k} f_k^{ji} & \text{if } k \text{ previously used} \\ \gamma f_{k^{new}}^{ji} & \text{if } k = k^{new} \end{cases}
$$

2.2.2 Sampling k

The dish assignments to tables in every document is governed by the distribution

$$
p(k_{jt} = k | \mathbf{t}, \mathbf{k}^{-jt}) \propto \begin{cases} m_{.k} f_k^{jt} & \text{if } k \text{ previously used} \\ \gamma f_{k^{new}}^{jt} & \text{if } k = k^{new} \end{cases}
$$

Note that while this equation looks similar to the equation for sampling a dish for a new table, it is important to note that the superscript is different reflecting the need to remove the entire table while computing the counts.

2.3 From unsupervised to semi-supervised

In this section, we explore a mechanism to inject human knowledge into the inference procedure to influence the development of topics during the MCMC procedure. Human input into the process consists of a set of topic prototypes $\mathcal{S} = \{s_1, s_2, \ldots\}$. Each element in the set s_i is a pair $< (s_{w_1}, s_{w_2}, ...), s_t >$ where $s_{w_1}, s_{w_2}, \ldots$ are words representative of a topic that a human thinks is present in the corpus. Similarly s_t is the approximate timestamp around which the topic is centered. The user is not required to provide both the list of words and a timestamp. Either of the two components may be missing in the topic prototype.

Incorporating human input of the form described is relatively straightforward in the MCMC sampling technique described earlier. For every topic s_i specified by the user, a corresponding pseudo-document is generated which contains only the top words specified for the topic and is stamped with the timestamp from the prototype. For these pseudo documents, only one table is assumed to be sampled and all the words in the document are assigned to this table. This single table is, in turn, assigned to a dish that is the counterpart of the prototype. During the MCMC sampling, the dish assignment and table sampling steps for the words in the pseudo documents are skipped and are instead set to a constant value as described. This allows the pseudo documents to contribute to the topic statistics and influence the sampling steps for regular documents while not changing membership themselves.

3 Datasets

We use two datasets to evaluate the HOT model. The NIPS dataset[1] consists of 1740 papers published in proceedings of NIPS 1-17. For our experiments, we use the abstracts of the papers. After stop word removal and elimination of words that occur fewer than 5 times, we obtain a dictionary of 8869 words.

We also use a corpus of blog posts focusing on American politics collected by Yano et al. [12]. The posts in the corpus span from Nov 2007 to the middle of 2008. The corpus was split into two subsets based on the political leanings of the blogs. The liberal blog corpus consists of posts from Daily Kos (http://www.dailykos.com) and Carpetbagger (http://www.thecarpetbaggerreport.com). The conservative blog corpus consists of posts from Red State (http://www.redstate.com) and Right Wing News(http://rightwingnews.com). As in the NIPS datasets, stopword and infrequent word elimination is performed. After processing the data, the liberal corpus contained 2419 documents with a vocabulary of 9,203 words and the conservative subset contained 2814 documents with a vocabulary of 8,934 words.

4 Results

In all the experiments below, we use five fold cross validation. The hyperparameters are sampled once for every experiment and the same hyperparameters are used for each fold. In experiments involving TOT and LDA, the smoothing parameters used for the topic specific multinomials are the same as in HDP and HOT. HOT and HDP require the setting of the concentration hyperparameters γ and α. We sample $\gamma \sim \text{Gamma}(1, 0.1)$ and $\alpha \sim \text{Gamma}(1, 1)$. Sample topics obtained after 1000 iterations of the MCMC sampling procedure are shown in Table 1. The graphs in the tables show the Gaussian distributions associated with the topics. Figure 2 shows the convergence properties of the MCMC sampling technique detailed in the previous section.

4.1 Evaluation

We evaluate the model in two ways - **document perplexity** and **time prediction**. Perplexity is given by $\exp\left(-\frac{1}{\text{\#Words in corpus}} \prod_{d \in docs} \prod_{w \in d} \log\left(p(w|\text{training set})\right)\right)$ and serves as an indicator of the quality of the fit of the model to the data provided. Figure 4 shows that LDA performs better than TOT on this metric for every cluster size. This is due to the fact that TOT is better able to model the document by using the timestamps of the documents. We further notice that HDP achieves nearly

[1] Obtained from Sam Roweis' webpage http://www.cs.toronto.edu/∼roweis/data.html

NIPS		Conservative		Liberal	
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2
recognition	likelihood	people	american	obama	carbon
character	density	american	iran	mccain	tax
digits	gaussian	faith	country	iowa	transit
character	parameter	god	iraq	delegates	rail
image	log	jesus	oil	state	oil
hand	em	christian	bush	nomination	fuel

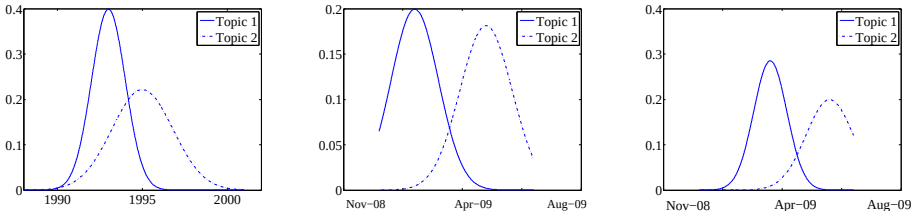


Table 1: Examples of topics obtained

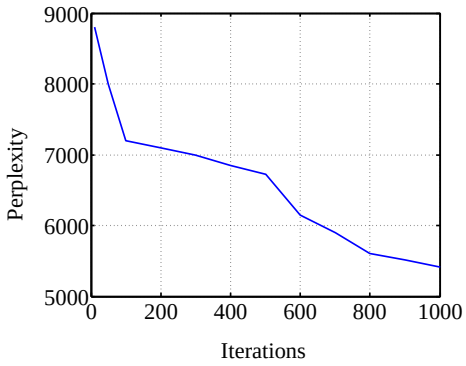


Figure 2: MCMC sampling convergence

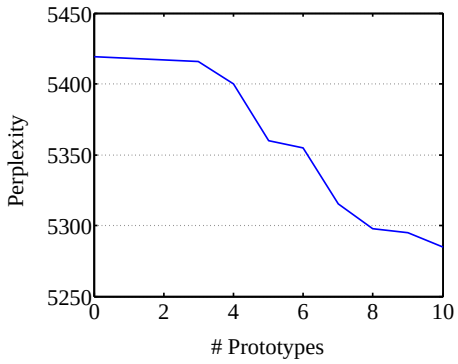


Figure 3: HOT with human guidance

the best perplexity achieved by LDA indicating that the hierarchical Dirichlet process is indeed able to choose the optimal number of topics automatically. Similarly HOT performs as well as the best TOT.

Time prediction is the task of predicting the timestamp of a previously unseen document. The predicted timestamp is evaluated against the actual timestamp of the document using squared error and accuracy is measured using the root mean squared error over the unseen corpus. The predicted timestamp for a document is obtained by first running MCMC sampling (without using the time components) to get the posterior topic probabilities which are subsequently used to get a weighted average of the means of the topic specific timestamp distributions. Figure 4 shows the root means squared error of TOT and HOT. Similar to document perplexity, we see that HOT gets a time prediction accuracy that is as good as the best TOT accuracy.

Figure 3 shows results from the NIPS dataset when human annotation is used to influence topic modeling. 3 to 10 topics were manually constructed from an inspection of the dataset and rough timestamps were assigned to the topic prototypes. Examples of topic prototypes provided are shown in Table 2. We can see from the graph that the document perplexity on unseen documents goes down as an increasing number of topic prototypes are provided. It should be noted that the decrease in perplexity is contingent on the quality of the topic prototypes provided.

5 Conclusion

In this work, we proposed a generative nonparametric topic model for modeling corpora of documents with timestamps. A Markov chain Monte Carlo sampling technique based on the Chinese

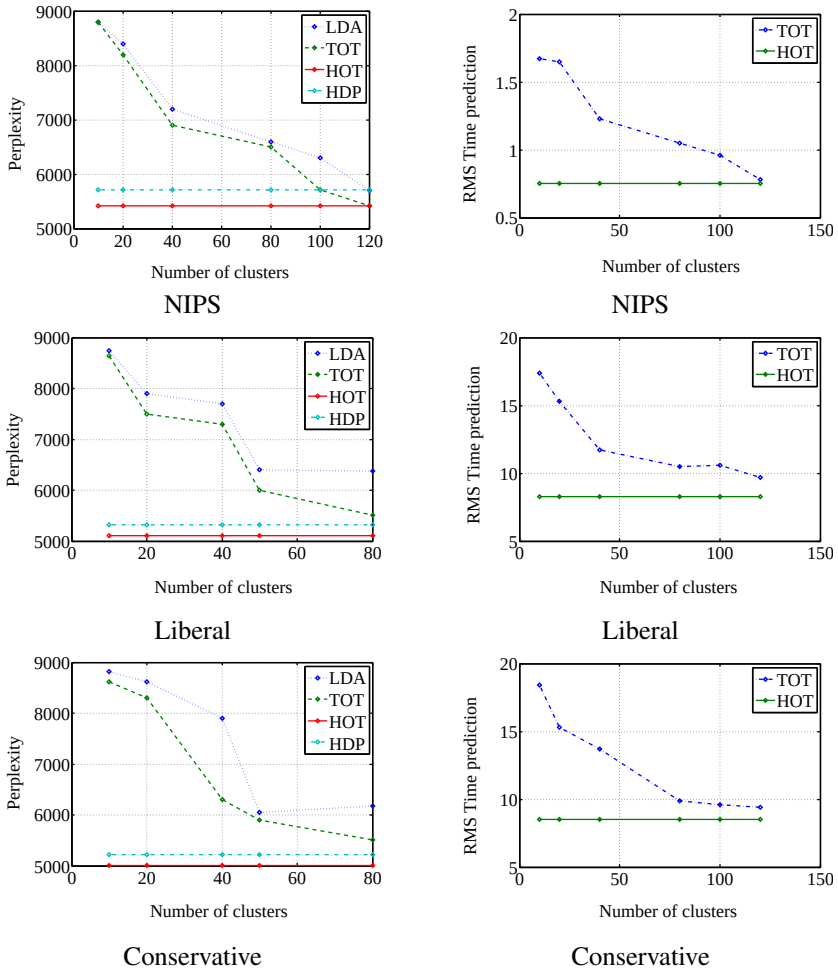


Figure 4: Perplexity and Time Prediction

speech, recognition, hmm, sequence, markov	1995
kernel, svm, support, margin, vectors	1999
bayesian, gaussian, parametric, priors, informative	-notimestamp-

Table 2: Topic prototypes

restaurant franchise view of hierarchical Dirichlet processes was presented to perform inference and obtain posteriors from the model. In experiments on the NIPS and blog datasets, the HOT model proposed showed better document perplexity than the HDP model which does not model time explicitly. HOT also performed better than TOT in the task of time prediction in unseen documents. We also presented a simple method to provide user input and showed how the sampling algorithm can be modified to accomodate user input. These results suggest that nonparametric topic models provide a principled way to tackle the problem of choosing the number of clusters and is an useful tool to analyze corpora with structures that vary over time.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[2] David M. Blei and Michael I. Jordan. Variational methods for the dirichlet process. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 12,

New York, NY, USA, 2004. ACM.

[3] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, New York, NY, USA, 2006. ACM.

[4] Ramesh M. Nallapati, Willian W. Cohen, Susan Ditmore, John D. Lafferty, and Kin Ung. Multiscale topic tomography. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–529, New York, NY, USA, 2007. ACM.

[5] Y. W. Teh. Dirichlet processes. Submitted to Encyclopedia of Machine Learning, 2007.

[6] C. Rasmussen. The infinite gaussian mixture model, 2000.

[7] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2003.

[8] Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *To appear in Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2009.

[9] Radford M. Neal and Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[10] David M. Blei and Jon D. Mcauliffe. Supervised topic models. 2007.

[11] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 121–130, New York, NY, USA, 2008. ACM.

[12] William W. Cohen Tae Yano and Noah A. Smith. Predicting response to political blog posts with topic models. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 2009.