A Co-evolving Correlated Dynamic Topic Model for Time Series Clickstream and Purchase Data

Hayden Luse, Alan Montgomery

16 May 2017

Abstract

Background. As an unprecedented amount of data about individual internet users has become available it is now possible to combine multiple channels of data to better profile users. We have clickstream data, online purchases and demographics purchases which we wish to use to understand behavior. Specifically we wish to user purchases to predict browsing behavior and vice versa. For example, a consumer's online browsing may help predict a trip before it begins, but then change drastically once it begins. A common analogy to clickstream and credit card purchase datasets is NLP and text mining however these approaches introduce several unjustified assumptions such as that a user purchasing a certain good in the past increases the probability that they will buy it in the future when exactly the opposite may well be true. Additionally many of these topic modelling approaches to the problem do not incorporate the time series nature of the problem.

Aim. In this paper we present a model which deals appropriately with all three types of data and within a single model, allowing a complete user description to be generated simultaneously instead of being cobbled together from models unaffected by the other datasets.

Data. In this paper we analyze the Comscore clickstream dataset containing demographic information, clickstream data and credit card purchase data for a population of users collected over 5 years.

Methods. In this paper we use a graphical model implementing a scalable inference scheme for a Correlated Dynamic Topic Model incorporating clickstream metadata. This model accounts for the bursty nature of clickstream data as well as the shifting of consumer behavior over a longer time horizon.

Results. Topics discovered using this model have greater purity than those of Latent Dirichlet Allocation (LDA) or the Correlated Topic Model (CTM) and also bursty behavior more accurately. The discovered topics are 11% more predictive of future purchasing behavior when used as features of a Random Forest than those of LDA or CTM.

Conclusions. This model appears to allow forward prediction of consumer purchasing categories. It appropriately models time series based secondary determiners of purchasing behavior such as previous purchases. Using this approach, the accuracy of internet ad targeting can be significantly improved.

Clickstream, Polya-Gamma, Scalable, Correlated Topic Model

1 Introduction

Businesses want to segment their customers holistically using online and offline behaviors. The challenge is that they rarely have data that tracks individual customers across all channels. For example, online retailers may have rich behavior about browsing and purchase at their individual website, but not other websites. Marketing researchers may have extensive knowledge about viewing across websites, but may only have information about online purchases. Our research shows how evolving topic models can be used to profile consumer behavior in each channel and then probabilistically relate these topics across channels. Specifically we predict the sequence of domains viewed or products purchased by relating these choices to latent, dynamic topics. Our approach accounts for the changing composition of factors within a channel, but also the way they relate to each other. For example, a topic like travel may appear in web browsing and then subsequently a related factor may temporarily appear in purchasing a month later due to vacation purchases. We apply our model to web browsing and online purchases from a ComScore clickstream panel of roughly 50,000 consumers with 300,000 purchases and 150 million viewings. The goal of this paper is to provide a methodology which fuses data about consumer behavior from many sources to allow improved consumer profiling for optimizing media planning or customer segmentation. The proposed model accounts for both correlated topics and temporal drift of topics and develops an efficient partially collapsed Gibbs sampler for model estimation.

2 Problem Statement

The problem being solved is to create a model that can be inferred efficiently but which also models data correlation and evolution over time. It is non-trivial to design a topic model that can appropriately identify the topics existing in each subset, model correlations between them at a user by user level, and model the changes in these factors between browsing and purchasing sessions. Further, inference for topic models addressing similar problems have been intractable at scale. This paper proposes a hierarchical correlated dynamic topic model with scalable Polya-Gamma based partially collapsed Gibbs sampler inference[Polson et al., 2013].

Related Work

Clickstream topic modeling is a fundamentally similar problem to topic modeling for general Natural Language Processing tasks[Mobasher, 2007]. The most popular starting point for topic models is Latent Dirichlet Allocation [Blei et al., 2003] and over the years this model has been expanded to represent more complicated features of language. In this paper we would like to represent topic correlation as in the Correlated Topic Model [Blei and Lafferty, 2007] and topic evolution over time as in the Dynamic Topic Model [Blei and Lafferty, 2006]. There have been several recent improvements which have allowed these models, but unfortunately the proposed methods do not scale well for our problem. A new method for dealing the non-conjugacy of the Correlated Topic Model is to use the augmented Polya-Gamma sampler [Polson et al., 2013, Chen et al., 2013]. Advances have also been made in taking advantage of the topic sparsity of the data have also been made [Ahmed et al., 2012] allowing the Gibbs Sampler inference to run at a lower computational complexity. There has also been a model designed that models multiple correlated text streams evolving over time

[Hong et al., 2011], however the authors did not allow for correlation across time or follow a fully Bayesian inference approach.

Acknowledgements

This research was sponsored by the Adobe Data Science Research Awards Program. We thank our colleagues from the Carnegie Mellon Machine Learning Department, especially who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions in this paper.

Data Description

We consider a Comscore consumer panel that tracks online browsing and purchasing for a representative sample of web users.[1]. The dataset has two parts: the first is top-level domains visited by a panelist and the second is online product purchases made by the panelists. The two datasets are anonymized, however panelists can be matched between them so that we can relate web browsing and online purchasing to each other. The data includes approximately 50,000 users with 250,000 purchases and 150 million viewings during the period between January 2011 and December 2012. This is a rich source of information to learn about customers.

Description	Number of Observations
Users	55,215
Page Views	338,287,825
Unique Domains	1,393,391
Sessions	2,353,338
Transactions	252,016

Table 1: Data Volume Description

The web browsing data includes top-level domains visited, the duration of the visit, the number of pages viewed within the top level domain, and the date and time of visit. The purchase transactions have a web browsing session ID, a date and time stamp of the transaction, the product name code, product category, quantity, price and the total price of the basket of goods purchased at one time.

The demographic information available about each panelist includes the head of household highest level of education, the head of household age, household income, household size, racial background, country of origin, number of children, internet connection speed, census region, and zip code. The panelists have a median household size of 3. 11% are originally from Hispanic countries.

Descriptive statistics about demographics are given in Table 2, 3 and 4. The panelists given minimal compensation to participate in the panel, for example, free virus protection software in exchange for allowing the monitoring of their web browsing and purchase data. There may be a bias in our panelists, since those that especially value their privacy may

[1]The Comscore data is available to all Carnegie Mellon University faculty and students through subscription http://search.library.cmu.edu/link/http://wrds-web.wharton.upenn.edu/wrds/connect

Oldest in Household	Percentage
18-20	4.2%
21-24	6.0%
25-29	7.8%
30-34	9.2%
35-39	9.4%
40-44	11.6%
45-49	12.8%
50-54	12.1%
55-59	9.0%
60-64	6.8%
65 and over	11.3%

Table 2: Oldest in Household Proportions

Household Income	Percentage
0-15k	13.2%
15k-24.9k	7.3%
25k-34.9k	10.0%
35k-49.9k	15.2%
50k-74.9k	26.2%
75k-99.9k	14%
100k and over	14.2%

Table 3: Household Income Proportions

Racial Background	Percentage
White	59.0%
Black	23.1%
Asian	3.9%
Other	14.3%

Table 4: Racial Background Proportions

not participate, but we see many sensitive activities which suggests that our viewers are representative of the population. An additional confounding factor is the inability to natively distinguish between multiple users of one machine. The following is a sample of the raw data. The first table is a single top level domain visit from a session and the second is a single purchase made during that session.

DOMAIN ID	39635416187341400000
MACHINE ID	9878417
SITE SESSION ID	3381629620292
DOMAIN NAME	bloomingdales.com
EVENT DATE	18700
REF DOMAIN NAME	yahoo.com
DURATION	1.4099121094
PAGES VIEWED	5
EVENT TIME	12:34:56

Figure 1: Raw data for a single top-level domain visit

DOMAIN ID	10902857555607900000
MACHINE ID	9878417
SITE SESSION ID	65477169451232
DOMAIN NAME	sears.com
EVENT DATE	18700
EVENT TIME	12:34:56
PROD NAME	GENUINE GRIP WOMEN SLIP-RESISTANT JOGGER WORK SHOES 1110 BLACK LEATHER
PROD CATEGORY ID	2
PROD QTY	2
PROD TOTPRICE	29.98
BASKET TOT	29.98

Figure 2: Raw data from a single purchase

3 Methods

3.1 Method

The method by which we analyzed the data was inspired by the Correlated Topic Model [Blei and Lafferty, 2007] and the Dynamic Topic Model [Blei and Lafferty, 2006] as well as recent developments in the training of both. The strategy was to design a graphical model combining the ability to model topic correlations present in the Correlated Topic Model and the ability to model topic proportions changing over time of the Dynamic Topic Model and to use this to represent the session to session changes in user browsing and purchasing behavior. We state the model in Figure 3.

The construction of the model begins with the CTM. If the nodes labeled U and U_t are removed this model is completely equivalent. The inclusion of node U adds a hierarchical dimension to the model. This allows us to model the differences in overall purchase and browsing topic distribution between two users. By including this node, we are able to describe individual users in a more straight-forward manner and with greater granularity. These nodes represent multivariate normal distributions so they are still able to model inter-topic correlations. Next we add the nodes labeled U_t. As these nodes are connected in series they allow for the representation of time series effects for an individual user. As

these nodes are once again multivariate normal, they can represent inter-topic correlations but also, importantly, they also model topic evolution as a gaussian random walk allowing the co-evolution of topics within the dataset to be explicitly modeled.

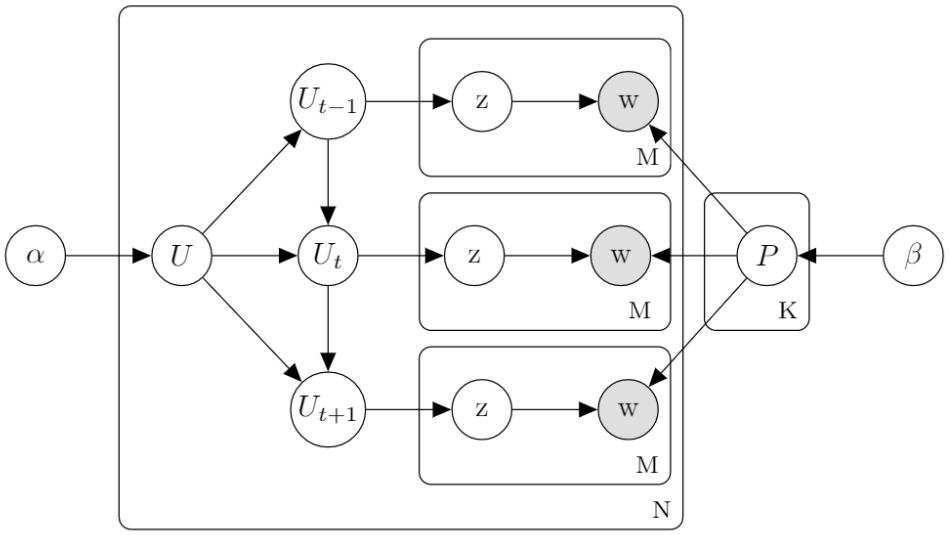

Figure 3: The plate notation graph structure of the model used in this paper where α is the global parameters Σ and μ

3.1.1 Sampling α

In this model, we first sample α by

$$\Sigma = IW(\kappa, W)$$

$$\mu = \mathcal{N}(\mu_0, \frac{\Sigma}{\rho})$$

Where, IW is the Inverse Wishart distribution. The hyperparameters were initialized to $\kappa = 100$, $\rho = 100$, $\mu_0 = 0$, $\delta = .8$ and W is the number of topics. in our experiments. The other unobserved nodes of the graph are sampled as follows.

$$U \sim \mathcal{N}(\mu, \Sigma)$$

$$U_t \sim \mathcal{N}(\delta U + (1-\delta)U_{t-1}, \Sigma)$$

$$z_{pt} \sim Mult(\pi(U_t))$$

$$P \sim Dir(\beta)$$

Where π is the softmax function and β is a vector of ones the length of the number of topics.

3.1.2 Sampling U_t

Sampling of this model using Variational Inference would be quite slow and not asymptotically correct, but recent advances have allowed much more scalable sampling of this model by using the Polya-Gamma distribution to derive an analytic solution for the conditional distribution of the U_t terms. The first step is to note that the likelihood of U_{nt}^k given $U_{\neg nt}^k$, the value of the k-th topic for the t-th time step for the n-th user, is through the Scale Mixture Representation [Chen et al., 2013] expressable as

$$\frac{e^{\rho_{nt}^k} C_{nt}^k}{(1+e^{\rho_{nt}^k})^{N_{nt}}} = \frac{1}{2^{N_{nt}}} * e^{\kappa_{nt}^k \rho_{nt}^k} \int_0^\infty e^{\frac{-\lambda_{nt}^k (\rho_{nt}^k)^2}{2}} p(\lambda_{nt}^k | N_{nt}, 0) d\lambda_{nt}^k$$

Where $\kappa_{nt}^k = C_{nt}^k - N_{nt}/2$, C_{nt}^k is the number of terms assigned to topic k for time step nt and N_{nt} is the number of words in the vocabulary of that time step. From here we have

$$\begin{gathered}
U_{nt} \sim \mathcal{N}(\gamma_{nt}^k, (\tau_{nt}^k)^2) \\
\gamma_{nt}^k = (\tau_{nt}^k)^2)(\sigma_k^{-2} \mu_{nt}^k + \kappa_{nt}^k + \lambda_{nt}^k \zeta_{nt}^k)^2 \\
(\tau_{nt}^k)^2) = (\sigma_k^{-2} + \lambda_{nt}^k)^{-1} \\
\sigma_k^2 = \wedge\wedge^{-1} \\
\wedge = \Sigma^{-1} \\
\zeta_{nt}^k = \log(\sum_{j \neq k} e^{U_{nt}^j}) \\
\mu_{nt}^k = \delta U + (1-\delta) U_{t-1} \\
\lambda_{nt}^k \sim \mathcal{PG}(N_{nt}, \rho_{nt}^k)
\end{gathered}$$

Since N_{nt} can be large, we use the fast, approximate sampler developed in Chen et al., 2013. In this sampler,

$$\begin{gathered}
y \sim \mathcal{PG}(n, \rho) \\
z \sim \mathcal{PG}(1, \rho) \\
y \approx \sqrt{Var(y)/Var(z)}(z - \mathbb{E}[z]) + \mathbb{E}[y] \\
\mathbb{E}[y] = \frac{n}{2\rho} tanh(\rho/2) \\
\frac{Var(y)}{Var(z)} = \frac{1}{n}
\end{gathered}$$

It has been shown that this is a very close approximation and reduces the time complexity of the sampling from $O(n)$ to $O(1)$.

3.1.3 Sampling U

Sampling U is simple as it is simply the Maximum A Posteriori estimate of a normal distribution.

3.1.4 Sampling z_{nt}^k

Another way to speed up the sampling process is to take advantage of the topic sparsity within a given browsing or purchasing session. We can achieve this by performing sparsity-aware fast sampling[Yao et al., 2009, Ahmed et al., 2012]. Specifically,

$$A_k = \frac{C_{k,\neg nt}^{w_n t}}{\sum_{j=1}^{V} C_{k,\neg nt}^{j} + \sum_{j=1}^{V} \beta_j}$$

$$B_k = \frac{C_{k,\neg nt}^{w_n t}}{\sum_{j=1}^{V} C_{k,\neg nt}^{j} + \sum_{j=1}^{V} \beta_j}$$

We sample z_{nt}^k from $Mult(B)$ with probability $p = \frac{\sum B_k}{\sum A_k + \sum B_k}$ and from $Mult(A)$ with probability $1 - p$.

3.1.5 Complexity

To reduce model update complexity we use a sparse version of the LDA topic update. Where K is the number of topics and s(K) is the average number of nonzero topics and S is the number of sub-burn in iterations, This reduces complexity to

$$O(N_d s(K) + K^2 + SK + N_U)$$

3.2 Design

In this experiment we follow the established path of other topic modeling papers. After fitting the model to the data, we first qualitatively analyze the topics discovered as generative models are only incompletely described by quantitative statistics. Additionally, we report the likelihood and perplexity of our model over time compared to other baseline models. Further, we attempt to predict future purchasing topics on the basis of past browsing and purchasing behavior by truncating the time series and report the accuracy, precision and recall. While it is difficult sometimes to evaluate the relative quality of generative models based on fundamentally different assumptions, we believe this suite of analyses should give a fair picture of the success of the proposed model.

4 Results

4.1 Analysis

We analyzed the results of our model's application in both qualitative and quantitative ways. Qualitatively, we analyze the purity of the discovered topics themselves, the correlation substructure of these topics, and the evolution of topics over time. Quantitatively we test whether topics discovered with our model are more predictive of future purchasing than those discovered by baseline models. We also test whether user similarity under our model is more predictive of similar purchasing behavior than under other models. In both cases, we compared the results of our model to results from LDA and the CTM.

4.2 Results

In the following sections, the baseline LDA and CTM comparisons were two separate models, both with 50 topics, one estimated on the browsing data and one estimated on the purchasing data. Our model was estimated on the full dataset with 50 browsing topics and 50 purchasing topics.

4.2.1 Qualitative Analysis

Our model appears to be successful at modeling the time series aspect of the clickstream data. Many times, purchase terms or URLs are correctly assigned to topics they would otherwise have had a very low probability of appearing in. This was most helpful with regards to book purchasing behavior, as part of the nature of book purchase terms is the possibility of any word in the language appearing in a title. Despite this, our model consistently applied words such as "laser" and "valentine" to topics related to book purchasing when appropriate based on the last session's topic distribution and to other categories when not. This behavior is not possible without the dynamic aspects of our model and led to much less precise topic assignment in this scenario by the other two models.

A representative example of our model being more predictive of purchasing behavior is shown in figures 2, 3 and 4. These graphs represent the actual and predicted purchasing behavior of a randomly selected user over 3 sequential browsing and purchasing sessions. While it is not a perfect match, in each case our model more closely represents actual purchasing behavior and most importantly evolves over time to more accurately predict the actual behavior.

Additionally, many of the correlations between browsing and purchasing topics accord with reasonable expectations. Browsing of travel related websites was correlated with future travel related purchases, however our model additionally decreased the likelihood of travel related purchases after they had been made. Our model also identified a link between unhealthy eating habits as evidenced by a large proportion of delivery pizza or chicken wings and future browsing of dieting related websites such as weightwatchers.com.

4.2.2 Quantitative Analysis

The first experiment run was to determine the average similarity of user purchasing behavior given similar user browsing behavior. For this, for each user the 3 most similar users based on cosine distance of their browsing topic distributions the cosine distance of their purchasing topic distributions was computed. Over the corpus, our model outperformed on this metric by 14% over LDA and 9% over the CTM as shown in Table 1.

Model	Average Similarity	Standard Deviation
LDA	.818	.06
CTM	.856	.09
CCDTM	.933	.04

Table 5: The cosine similarity and standard deviation thereof between the purchasing topics of a user and their three nearest neighbors by cosine similarity of their browsing topics

The second experiment was performed by choosing random points in time and estimating the model with only access to data from before that time. A Random Forest [Kam, 1995]

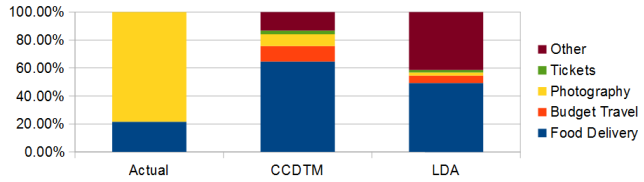

Figure 4: Actual and Predicted purchasing topic prevalence for time period 1

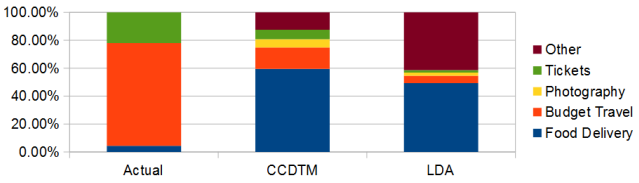

Figure 5: Actual and Predicted purchasing topic prevalence for time period 2

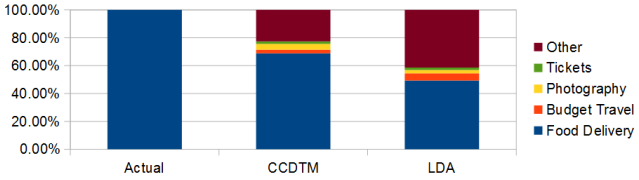

Figure 6: Actual and Predicted purchasing topic prevalence for time period 3

was then trained to predict the most common purchase topic. The likelihood of the observed maximum purchasing topic for each model was then evaluated. Our model outperformed the CTM by 11% and LDA by 18% as shown in Table 2.

Model	Likelihood
CCDTM	.460
CTM	.414
LDA	.390

Table 6: The likelihood of the observed future most common purchasing topic for users based on a random forest trained on previous user purchasing and browsing topics

Next, we examined the average perplexity of the models. When examining the perplexity of each model, our model outperformed by 5.7% over the CTM and 33.6% over LDA. However, our model took several times longer to train than a normal Gibbs Sampler based CTM. This is expected for a more complicated model and the training was manageable if not ideal.

Model	Perplexity	Running Time
LDA	2140	342s
CTM	1508	7254s
CCDTM	1421	21189s

Table 7: Perplexity and Run Time of each model on the Comscore Clickstream Dataset

As a next experiment, we use the same random forest scheme as before. We calculate the perplexity for LDA, CTM and our model predicting purchase data from only clickstream data and vice versa using different length data histories. Specifically, we train a model each using only the last click or purchase, using the last session, using the last 10 sessions, and using all available data.

History	LDA	CTM	CCDTM
1 Click	8640	8222	8131
1 Session	3027	1545	1668
10 Sessions	2926	1288	951
All Data Cat	2712	1152	874

Table 8: Perplexity when predicting purchase topics from only clickstream data

History	LDA	CTM	CCDTM
1 Purchase	9347	9132	8928
1 Session	8027	7285	7324
10 Sessions	7874	6238	5321
All Data Cat	7743	6156	5189

Table 9: Perplexity when predicting browsing topics from only purchase data

5 Discussion

This analysis of the Comscore Clickstream dataset appears to validate the initial hypothesis that a holistic model trained on both data streams would outperform models which did not allow them to interact. Our model appears significantly more predictive of purchase topics given browsing topics than the other models which implies that this approach was more well suited to this type of data.

The Dynamic Co-evolving Correlated Topic Model designed in this paper appears to have been successful at modeling the complex interactions between browsing and purchasing data while retaining relatively efficient inference. We believe that this is the first multi-stream correlated dynamic topic model based on an augmented polya-gamma sampler. As multi-stream textual or more generally categorical data is common in many fields, most famously Natural Language Processing, we believe that this model has a natural extension to other

areas of research. More specifically, this approach using Gibbs Sampler based inference appears to be more easily scalable and parallelizable [Smola and Narayanamurthy, 2010] than previous models with these features [Hong et al., 2011, Song et al., 2008].

We believe that this model can be useful for a variety of tasks. It appears to be very well suited for predicting user purchase similarity based on browsing behavior, which is a particularly interesting quality for many content providers. Using predicted similarities from this model, it should be possible to provide efficient descriptions of a service's users to help better tailor that service to their needs.

6 Limitations

In this paper, we were limited in our analysis of the effect of the number of topics on the performance of the model. While we have theoretical results with regards to the time complexity of the Gibbs Sampler we were not able to perform extensive experiments with changing topic number. In the topic models which inspired our model [Chen et al., 2013, Yao et al., 2009], an increased number of topics led to a logarithmically increasing runtime, and while in theory our model complexity should be a constant offset of the aforementioned logarithmically increasing runtime, we have not performed the necessary experiments to validate this due to time constraints.

In our dataset, we were limited to a time period ending 5 years ago, and we cannot ignore the possibility that the underlying dynamics of this type of data have changed although we believe that is unlikely. Further research would be well served by using more recent data if available.

7 Conclusion

In this paper we have described and applied a Co-evolving Correlated Dynamic Topic Model holistically to clickstream and purchase data. We have achieved performance superior to common existing models and have been able to capture time varying topic proportions and inter-topic correlation. Using this approach, browsing behavior of users can be effectively used to predict purchase behavior similarity allowing content providers to more accurately and efficiently describe and predict their user-base. In addition, the efficient inference methods used in this paper could also be applied to single stream or Natural Language Processing data with ease and may be a more effective method of describing some data in those regimes. For example, a collection text posts by social media users have the same user and user session structure so it is a natural extension of this work to apply this model to that type of data.

The Co-evolving Correlated Dynamic Topic Model represents a successful fusion of several disparate topic modeling techniques and accounts for temporal and correlative effects in the data as well as exploiting its sparsity. We have implemented a more efficient inference method than the variational inference that would otherwise have been required. It is our hope that extensions of this model could be used in other areas or on larger, web-scale datasets in the future.

References

[Ahmed et al., 2012] Ahmed, A., Aly, M., Gonzalez, J., Narayanamurthy, S., and Smola, A. (2012). Scalable inference in latent variable models. In *International conference on Web search and data mining (WSDM)*, volume 51, pages 1257–1264.

[Blei and Lafferty, 2006] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

[Blei and Lafferty, 2007] Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Chen et al., 2013] Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453.

[Hong et al., 2011] Hong, L., Dom, B., Gurumurthy, S., and Tsioutsiouliklis, K. (2011). A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 832–840. ACM.

[Kam, 1995] Kam, H. T. (1995). Random decision forest. In *Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition, Montreal, Canada, August*, pages 14–18.

[Mobasher, 2007] Mobasher, B. (2007). Data mining for web personalization. In *The adaptive web*, pages 90–135. Springer.

[Polson et al., 2013] Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using polya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

[Smola and Narayanamurthy, 2010] Smola, A. and Narayanamurthy, S. (2010). An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710.

[Song et al., 2008] Song, Y., Zhang, L., and Giles, C. L. (2008). A non-parametric approach to pair-wise dynamic topic correlation detection. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 1031–1036. IEEE.

[Yao et al., 2009] Yao, L., Mimno, D., and McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM.