LETTER

A Most Resource-Consuming Disease Estimation Method from Electronic Claim Data Based on Labeled LDA

Yasutaka HATAKEYAMA[†a)], **Takahiro OGAWA**[†], ***Members***, **Hironori IKEDA**[†], ***Nonmember***, ***and*** **Miki HASEYAMA**[†], ***Member***

SUMMARY In this paper, we propose a method to estimate the most resource-consuming disease from electronic claim data based on Labeled Latent Dirichlet Allocation (Labeled LDA). The proposed method models each electronic claim from its medical procedures as a mixture of resource-consuming diseases. Thus, the most resource-consuming disease can be automatically estimated by applying Labeled LDA to the electronic claim data. Although our method is composed of a simple scheme, this is the first trial for realizing estimation of the most resource-consuming disease.

key words: *most resource-consuming disease, electronic claim, labeled LDA*

1. Introduction

Health insurance claims (HICs) are prepared by providers such as hospitals for reimbursement of their services. In Japan, HIC records contain various types of information including information on the patient's name, date of birth, gender, health insurance qualification status, medical procedures, costs of medical procedures and diseases. In order to determine priorities in health policy, evaluation of the costs of medical procedures is necessary.

Because of this situation, online billing of HICs became mandatory in Japan in 2011*. In this paper, we denote the online bill of HICs as "electronic claim". Electronic claims also include histories of medical resource consumption, and several methods for analysis of electronic claims have been proposed [1], [2]. These approaches can identify the characteristics of diseases from electronic claim data by introducing medical knowledge. However, these approaches need not only manual classification of medical procedures and diseases but also advanced medical knowledge. Therefore, a fully automatic method for analysis of electronic claims that does not require manual classification or advanced medical knowledge is needed. Furthermore, not only the problem of identification of the characteristics of diseases but also several new problems have become important. One of the most important problems in recent years is estimation of the most resource-consuming diseases, since we have to determine priorities in health policy for reduction of healthcare costs.

Manuscript received June 4, 2015.
Manuscript revised October 27, 2015.
Manuscript publicized November 30, 2015.

[†]The authors are with Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: hatakeyama@lmd.ist.hokudai.ac.jp

DOI: 10.1587/transinf.2015EDL8128

In order to estimate the most resource-consuming diseases without any manual classification or advanced medical knowledge, an automatic estimation method based on relationships between diseases and medical procedures is needed.

In recent years, various topic estimation methods from document collections have been proposed [3], [4]. These approaches model each document from words in the document as a mixture of topics that represent the meaning of the documents. Interestingly, we have found that electronic claims, medical procedures and diseases can be regarded as documents, words, and topics, respectively. Furthermore, costs of diseases that are recorded in electronic claims are calculated on the basis of costs of the medical procedures. It can therefore be expected that estimation of the most resource-consuming disease becomes feasible on the basis of the costs of the medical procedures and the relationships between diseases and medical procedures obtained from the topic model.

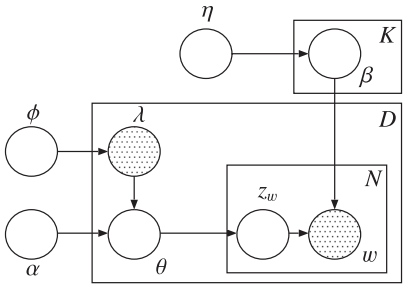
In this paper, a method to estimate the most resource-consuming disease from electronic claim data based on Labeled Latent Dirichlet Allocation (Labeled LDA) [4] is presented. The proposed method models each electronic claim from its medical procedures as a mixture of diseases. The most resource-consuming disease can be automatically estimated by applying Labeled LDA to the electronic claim data without any manual classification or advanced medical knowledge. Although our method consists of a simple scheme, this is the first trial for realizing estimation of the most resource-consuming disease.

This paper is organized as follows. As preliminaries, Labeled LDA is explained in Sect. 2. In Sect. 3, the proposed method for estimation of the most resource-consuming diseases is presented. In Sect. 4, results of experiments obtained by applying the proposed method to actual electronic claim data are shown to verify the effectiveness of the proposed method. Concluding remarks are given in Sect. 5.

2. Labeled LDA

In this section, we briefly explain Labeled LDA [4]. Labeled LDA is a probabilistic topic model that describes a process

*Ministry of Health, Labour and Welfare announced "Health Care Reform Outline" in 2006.



Labeling prior probability ϕ
Parameter of the Dirichlet topic prior α
Distributions over all topics θ
Distributions over all words β.
Topic presence/absence indicator λ
Parameter of Dirichlet Distribution η
Topics z_w
Words w

Fig. 1 Graphical model of Labeled LDA.

for generating a labeled document collection. Like Latent Dirichlet Allocation (LDA) [3], Labeled LDA models each document as a mixture of underlying topics, and each word in the documents is generated from one topic. Compared to LDA, Labeled LDA incorporates supervision by simply constraining the topic model to use only those topics that correspond to a document's observed label set. Figure 1 shows a graphical model of Labeled LDA.

Let each document $d \in \{1, \cdots, D\}$ be represented by a tuple consisting of a list of word indices $\boldsymbol{w}_d = (w_{d,1}, \cdots, w_{d,N_d})$ and a list of binary topic presence/absence indicators $\boldsymbol{\lambda}_d = (l_{d,1}, \cdots, l_{d,K})$, where each $w_{d,n} \in \{1, \cdots, V\}$ and each $l_{d,k} \in \{0, 1\}$. Note that N_d is the total number of words in the document d, V is the total number of unique words and K is the total number of unique labels in the dataset. Specifically, the document collection $\mathcal{D} = \{(\boldsymbol{w}_1, \boldsymbol{\lambda}_1), \cdots, (\boldsymbol{w}_D, \boldsymbol{\lambda}_D)\}$ is generated as shown in Fig. 2. Note that $\boldsymbol{L}_d = \{L_d(i, j)\}$ is defined as follows:

$$L_d(i, j) = \begin{cases} 1 & \text{if} \quad \lambda_d(i) = j \\ 0 & otherwise. \end{cases} \tag{2}$$

Note that $\lambda_d(i)$ is the i-th element of the document's label vector $\boldsymbol{\lambda}_d = \{k | \lambda_d(k) = 1\}$. In other words, the i-th row of L_d has an entry of 1 in column j if and only if the i-th document label $\lambda_d(i)$ is equal to the topic j.

In order to estimate parameters $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_{z_{d,n}}$, Collapsed Gibbs Sampling [5] is utilized in [4]. Note that $\boldsymbol{\theta}_d$ is the topic distribution of each document, and $\boldsymbol{\beta}_{z_{d,n}(=j)}$ is the word distribution of each topic. Here the labeled topic of the document collection is represented by $\boldsymbol{w} = \{w_{d,n}\}$. Furthermore, the labeled topic of each word in the document collection is respectively represented by $\boldsymbol{z} = \{z_{d,n}\}$. We utilize Collapsed Gibbs sampling for the training where the sampling probability for a topic for position i in a document d in Labeled LDA is given by:

(1) For each topic $k \in \{1, \ldots, \mathcal{K}\}$
  (a) Generate distributions over all words $\boldsymbol{\beta}_k = (\beta_{k,1}, \cdots, \beta_{k,V})^T$, which is a vector consisting of parameters of the multinomial distribution corresponding to the k-th topic from the Dirichlet distribution $Dir(\boldsymbol{\eta})$.
(2) For each document $d \in \{1, \ldots, \mathcal{D}\}$
  (a) For each topic $k \in \{1, \cdots, K\}$
    i. Generate document's labels $l_{d,k}$ by using $Bernoulli(\Phi_k)$ for each topic k with a labeling prior probability Φ_k.
  (b) Generate parameters of the Dirichlet distribution $\boldsymbol{\alpha}_d$ as the follow equation:

$$\boldsymbol{\alpha}_d = \boldsymbol{L}_d \times \boldsymbol{\alpha}. \tag{1}$$

  Note that $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^T$ is the parameter vector of the Dirichlet topic prior.
  (c) Generate distributions over all topics $\boldsymbol{\theta}_d$ from $Dir(\boldsymbol{\alpha}_d)$.
  (d) For each word $n \in \{1, \cdots, N_d\}$
    i. Generate topic $z_{d,n}$ from multinomial distributions $Multi(\boldsymbol{\theta}_d)$.
    ii. Generate word $w_{d,n}$ from multinomial distributions $Multi(\boldsymbol{\beta}_{z_{d,n}})$.

Fig. 2 Algorithm of Labeled LDA.

$$P(z_{d,n} = j | \boldsymbol{z}_{-(d,n)}, \boldsymbol{w}) \propto \frac{n^{w_{d,n}}_{-(d,n),j} + \eta_{w_{d,n}}}{\sum_{w'=1}^{V} n^{w'}_{-(d,n),j} + \eta_w} \times \frac{n^{d}_{-(d,n),j} + \alpha_j}{\sum_{j'=1}^{K} n^{d}_{-(d,n),j'} + \alpha_j} \quad \text{subject to} \quad j \in \boldsymbol{\lambda}_d. \tag{3}$$

Note that $z_{-(d,n)}$, $n^{w_{d,n}}_{-(d,n),j}$ and $n^{d}_{-(d,n),j}$ respectively represent $\boldsymbol{z}$ not including the current assignment $z_{d,n}$, counts of words $w_{d,n}$ in topic j not including the current assignment $z_{d,n}$ and count of generation of topic j in document d not including the current assignment $z_{d,n}$. Additionally we suppose $\eta_1 = \cdots = \eta_V$ and $\alpha_1 = \cdots = \alpha_K$ in the derivation of Eq. (3). By using the obtained z from Eq. (3), $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_{z_{d,n}(=j)}$ are respectively estimated as follows:

$$\theta_{d,j} = \frac{n^{d}_{j} + \alpha_j}{\sum_{j'=1}^{K} n^{d}_{j'} + \alpha_j}, \tag{4}$$

$$\beta_{j,w} = \frac{n^{w}_{j} + \eta_w}{\sum_{w'=1}^{V} n^{w'}_{j} + \eta_w}. \tag{5}$$

As shown in the above procedures, Labeled LDA models each document as a mixture of underlying topics. Furthermore, each word is generated from one topic.

3. Estimation of the Most Resources-Consuming Disease Based on Labeled LDA

In this section, we propose a method of estimating the most resources-consuming disease based on Labeled LDA. First,

we regard electronic claims, medical procedures that are described in the electronic claim and diseases as documents, words, and topics, respectively. As described above, DPC data† are one of electronic claim data. DPC data consist of two files that are "anonymized information obtained from clinical records" and "details of the medical procedures and treatment". In "anonymized information obtained from clinical records", diagnosed diseases, identification number, date of discharge, date of admission, etc. are described. Furthermore, medical procedures, their costs, their execution date, identification number, date of discharge, date of admission, etc. are described in "details of the medical procedures and treatments". An example of "details of the medical procedures and treatments" is shown in Table 4. In the proposed method, we associate "anonymized information obtained from clinical records" with "details of the medical procedures and treatments" by using identification number, date of discharge and date of admission. Therefore, the proposed method can estimate the most resource-consuming disease from the associated DPC data.

By applying Labeled LDA to these electronic claim data, relationships between the selected medical procedures and diseases are obtained. Furthermore, costs of the diseases that are described in electronic claims are calculated on the basis of costs of the medical procedures. Then the most resource-consuming disease can be estimated on the basis of the costs of medical procedures for each disease. The details are shown below. Let each electronic claim d be represented by a tuple consisting of a feature vector of medical procedures $\boldsymbol{w}_d = (w_{d,1}, \cdots, w_{d,N_d})$ and a list of binary disease presence/absence indicators $\boldsymbol{\lambda}_d = (l_{d,1}, \cdots, l_{d,K})$, where each $d \in \{1, \ldots, D\}$, each $w_{d,n} \in \{1, \cdots, V\}$ and each $l_{d,k} \in \{0, 1\}$. Furthermore, a feature vector of the described diseases in electronic claim d is $\boldsymbol{l}_d = \{k | l_{d,k} = 1\}$. Note that D is the total number of electronic claims, N_d is the total number of described medical procedures in the electronic claim d, K is the number of diseases, and V is the total number of medical procedures. Next, we apply Labeled LDA to the electronic claim class $\mathcal{D} = \{(\boldsymbol{w}_1, \boldsymbol{\lambda}_1), \cdots, (\boldsymbol{w}_D, \boldsymbol{\lambda}_D)\}$. Then probabilistic distributions over medical procedures $\boldsymbol{\beta}_k = (\beta_{k,1}, \cdots, \beta_{k,V})$ and probabilistic distributions $\boldsymbol{\theta}_d$ over all K diseases for each electronic claim d are obtained. In the proposed method, costs of diseases are calculated from the medical procedures and costs of these medical procedures. Specifically, the cost of each disease $k \in \boldsymbol{\lambda}_d$ is calculated by the following equation:

$$point^{disease}(d, k) = point(\boldsymbol{w}_d) I(k, \boldsymbol{w}_d). \tag{6}$$

Note that $point(\boldsymbol{w}_d)$ is a function that represents the costs of medical procedure $\boldsymbol{w}_d$, which is a medical bill obtained from the electronic claim. Furthermore, $I(k, \boldsymbol{w}_d)$ is an indicator function as follows:

$$I(k, \boldsymbol{w}_d) = \begin{cases} 1 & \text{if } \frac{P(k|\boldsymbol{w}_d, d)}{\max_{k' \in \boldsymbol{\lambda}_d} P(k'|\boldsymbol{w}_d, d)} \geq Th^I \\ 0 & otherwise, \end{cases} \tag{7}$$

where Th^I is a threshold value whose range is $[0, 1]$. Then $P(k|\boldsymbol{w}_d, d)$ is calculated from parameters $P(k|d)$ $(= \theta_{d,k})$ and $P(\boldsymbol{w}_d|k)$ $(= \beta_{k,w})$ that are obtained by Labeled LDA by the following equation:

$$P(k|\boldsymbol{w}_d, d) = \frac{P(k|d) P(\boldsymbol{w}_d|k)}{P(\boldsymbol{w}_d)}. \tag{8}$$

Note that $P(k|d)$ and $P(\boldsymbol{w}_d|k)$ in Eq. (8) correspond to $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_{z_{d,n}(=j)}$ in Eqs. (4) and (5), respectively. As shown in the above procedures, the most resource-consuming disease k in the electronic claim d can be estimated by using Eq. (6).

In the proposed method, relationships between resource-consuming diseases and medical procedures can be estimated. Thus, the proposed method can estimate the most resource-consuming disease based on the costs of medical procedures without any advanced medical knowledge.

4. Experimental Results

The effectiveness of the proposed method for estimating the most resource-consuming disease based on Labeled LDA is verified in this section. In the experiments, we utilized 69532 $(= D)$ Diagnosis Procedure Combination [6] (DPC) data that include 114 $(= K)$ kinds of diseases as electronic claims and 274848 $(= V)$ medical procedures. DPC data consist of combinations of some disease names and medical procedures. The most resource-consuming disease name is also manually described. Note that the most resource-consuming disease name is not necessary to apply the proposed method. This information is only utilized for quantitative evaluation and for conventional methods. Parameters of the proposed method that were used in experiments are shown in Table 1.

Experimental results are shown in Table 2 to verify the effectiveness of the proposed method. From Table 2, we can see that the proposed method can estimate the most resource-consuming disease by using the probability value and the cost of medical procedures. In the proposed method, not only posterior probability over each disease but also costs of medical procedures are considered. Therefore, the most resource-consuming disease can be correctly estimated.

Next, we quantitatively verify the effectiveness of the proposed method. We used recall, precision and F-value, which are defined as follows:

$$\text{Recall} = \frac{\text{Num. of correctly estimated electronic claims}}{\text{Num. of relevant electronic claims}}, \tag{9}$$

Table 1 Experimental parameters of the proposed method.

K	α_j $(j = 1. \cdots, K)$	η_w $(w = 1, \cdots, V)$	T	Th^I
161	$50/K$	0.1	300	0.1

†Details of DPC data are shown in http://www.mhlw.go.jp/topics/2011/09/tp0928-1.html (Japanese)

Table 2 Experimental results of the proposed method.

(a) Experimental results for which the most resource-consuming disease is "Crohn's disease".

Estimated most resource-consuming disease	Diseases described in electronic claim	Probability value ($= P(k\|\boldsymbol{w}_d, d)$)	Costs of disease (JPY) ($= point^{disease}(d,k)$)
Crohn's disease	Crohn's disease	0.33	¥2,570,800
	Esophagus, stomach, duodenum and other intestinal inflammation (other benign disease)	0.29	¥1,902,330
	Vestibular dysfunction	0.08	¥2,444,830

(b) Experimental results for which the most resource-consuming disease is "Bladder tumor".

Estimated most resource-consuming disease	Diseases described in electronic claim	Probability value ($= P(k\|\boldsymbol{w}_d, d)$)	Costs of disease (JPY) ($= point^{disease}(d,k)$)
Bladder tumor	Bladder tumor	0.37	¥3,868,320
	Type 2 diabetes (except for diabetic ketoacidosis)	0.52	¥2,489,618
	Sleep disorders	0.03	¥813,990
	Other malignancies	0.01	¥731,800

(c) Experimental results for which the most resource-consuming diseases are "Pneumonia, acute bronchitis and bronchiolitis".

Estimated most resource-consuming disease	Diseases described in electronic claim	Probability value ($= P(k\|\boldsymbol{w}_d, d)$)	Costs of disease (JPY) ($= point^{disease}(d,k)$)
Pneumonia, acute bronchitis and bronchiolitis	Pneumonia, acute bronchitis and bronchiolitis	0.22	¥193,380
	Type 2 diabetes (except for diabetic ketoacidosis)	0.36	¥184,050
	Chronic nephritic syndrome, interstitial nephritis and renal failure	0.02	¥155,120

$$\text{Precision} = \frac{\text{Num. of correctly estimated electronic claims}}{\text{Num. of all estimated electronic claims.}}, \tag{10}$$

$$\text{F-value} = \frac{2.0 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{11}$$

Experimental results obtained by using the proposed method and the following conventional methods are shown in Table 3.

Conventional method (i)

By using Support Vector machine [7] (SVM), one-against-one [8] multi-class classifiers were constructed for estimation of the most resource-consuming disease. In this conventional approach, by aligning the cost of each medical procedure, normalized feature vectors are extracted from each electronic claim. The most resource-consuming disease obtained from DPC data is assigned as the class label. Furthermore, we use RBF kernel $\exp(-\gamma\|x - x'\|^2)$. Note that γ is 3.6×10^{-6}, and the cost parameter C is 1.

Conventional method (ii)

By using multinomial naive bayes, multi-class classifiers were constructed for estimation of the most resource-consuming disease. In the same the way as the above conventional method (i), normalized feature vectors are extracted from each medical procedure. Furthermore, the most resource-consuming disease is assigned as the class label.

From Table 3, we can see that the proposed method has higher recall, precision and F-value compared to the conventional methods. Compared to the results obtained by using the conventional methods, the proposed method can realize accurate estimation without a training dataset including the most resource-consuming disease data. The proposed method models each electronic claim from medical procedures in the electronic claim as a mixture of resource-consuming diseases. By using the obtained model, the proposed method can accurately estimate the most resource-consuming disease. Therefore, the proposed method is more appropriate than the conventional methods. In the original Labeled LDA, the topics from a document can be estimated with the topic probabilities. Additionally, the topic probabilities are calculated from all annotated topics of the document collection. On the other hand, a topic (disease) is already annotated in the DPC data in our method, and we do not know only the topic probabilities. Then we calculate only the topic (disease) probabilities of described diseases in the target DPC data, and the costs of medical procedures are also utilized in the proposed method. Therefore, the proposed method can realize higher performance than the original Labeled LDA.

Table 3 Experimental results of the proposed method and conventional methods.

	Our method	Conventional method (i)	Conventional method (ii)
Recall	84.3	60.0	29.5
Precision	88.4	85.9	79.1
F-value	85.2	70.7	43.0

5. Conclusion

In this paper, we propose a method to estimate the most

Table 4 Example of "details of the medical procedures and treatments".

Health industry number	Identification number	Date of discharge	Date of admission	Medical procedure number	Medical procedure name	Dosage amount	Itemized point	Yen/Point flag	Fee-for-service reimbursement point	The number of the medical procedures	Health insurance number	Execution date
010116258	0000000095	20100331	20100329	000	Catract surgery	0	0	0	0	1	390110	20100330
010116258	0000000095	20100331	20100329	001	Catract surgery	0	12100	0	12100	1	NULL	20100330
010116258	0000000095	20100331	20100329	002	Cefamezin α for Injection	1	797	1	0	1	NULL	20100330
010116258	0000000095	20100331	20100329	003	Ecolicin ophthalmic ointment	0.5	17.55	1	0	1	NULL	20100330
010116258	0000000095	20100331	20100329	004	Opegan hi 1% 0.85ml	1	8542.5	1	0	1	NULL	20100330
010116258	0000000095	20100331	20100329	005	Healon V0.6 Ophthalmic Viscoelastic Substance 2.3%	1	10932	1	0	1	NULL	20100330
010116258	0000000095	20100331	20100329	006	Isotonic Sodium Chloride Solution 100ml	2	184	1	0	1	NULL	20100330

resource-consuming disease from electronic claim data based on Labeled LDA. The proposed method models each electronic claim from medical procedures in the electronic claim as a mixture of resource-consuming diseases. By applying Labeled LDA to the electronic claim data, the most resource-consuming disease can be estimated without any advanced medical knowledge. Our experimental results verified the effectiveness of the proposed estimation method.

This is the first trial for realizing estimation of the most resource-consuming disease. Therefore, in order to improve estimation accuracy, we consider introduction of the other topic model methods such as [9] and [10]. Reference [9] can consider the power-law phenomenon of a word distribution which is known as Zipf's law in linguistics and presence of multiple topics. By using reference [9], we can consider Zipf's law. Reference [10] proposed probabilistic time series models. In the DPC data, execution date of the medical procedures are contained. By introducing reference [10] to our method, improvement of accuracy will be expected. These will be the subjects of subsequent studies.

Acknowledgments

In the experiments, we utilized DPC data that were provided by associate professor Kenji Fujimori, Division of Medical Management, Hokkaido University Hospital. This work was partly supported by Grant-in-Aid for Scientific Research (B) 25280036, Japan Society for the Promotion of Science (JSPS).

References

[1] S. Tanihara, "Assessment of text documentation accompanying uncoded diagnoses in computerized health insurance claims in Japan," Journal of Epidemiology, vol.25, no.3, pp.181–188, 2015.
[2] N. Shoda, H. Yasunaga, H. Horiguchi, S. Matsuda, K. Ohe, Y. Kadono, and S. Tanaka, "Risk factors affecting inhospital mortality after hip fracture: Retrospective analysis using the japanese diagnosis procedure combination database," BMJ Open, vol.2, no.3, e000416, 2012.
[3] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol.3, pp.993–1022, March 2003.
[4] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, "Labeled IDA: A supervised topic model for credit attribution in multi-labeled corpora," Proc. 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP '09, vol.1, pp.248–256, Stroudsburg, PA, USA, ACL, 2009.
[5] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. National Academy of Sciences of the United States of America, vol.101 Suppl. 1, pp.5228–5235, April 2004.
[6] S. Matsuda, K. Fujimori, and K. Fushimi, "Development of casemix based evaluation system in Japan," Asian Pacific Journal of Disease Management, vol.4, no.3, pp.55–66, 2010.
[7] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, New York, NY, USA, 1995.
[8] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw., vol.13, no.2, pp.415–425, March 2002.
[9] I. Sato and H. Nakagawa, "Topic models with power-law using Pitman-Yor process," Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pp.673–682, New York, NY, USA, ACM, 2010.
[10] D.M. Blei and J.D. Lafferty, "Dynamic topic models," Proc. 23rd International Conference on Machine Learning, ICML '06, pp.113–120, New York, NY, USA, ACM, 2006.