Coupled Hierarchical Dirichlet Process Mixtures for Simultaneous Clustering and Topic Modeling

Masamichi Shimosaka[1] (✉), Takeshi Tsukiji[2], Shoji Tominaga[2], and Kota Tsubouchi[3]

[1] Tokyo Institute of Technology, Tokyo, Japan
simosaka@miubiq.cs.titech.ac.jp
[2] The University of Tokyo, Tokyo, Japan
tsukiji@miubiq.cs.titech.ac.jp,tominaga@ics.t.u-tokyo.ac.jp
[3] Yahoo Japan Corporation, Tokyo, Japan
ktsubouc@yahoo-corp.jp

Abstract. We propose a nonparametric Bayesian mixture model that simultaneously optimizes the topic extraction and group clustering while allowing all topics to be shared by all clusters for grouped data. In addition, in order to enhance the computational efficiency on par with today's large-scale data, we formulate our model so that it can use a closed-form variational Bayesian method to approximately calculate the posterior distribution. Experimental results with corpus data show that our model has a better performance than existing models, achieving a 22% improvement against state-of-the-art model. Moreover, an experiment with location data from mobile phones shows that our model performs well in the field of big data analysis.

Keywords: Non-parametric Bayes · Clustering · Hierarchical model · Topic modeling

1 Introduction

In this paper, we focus on a nonparametric Bayesian model in which the complexity of data can be controlled by using a stochastic process such as the Dirichlet process (DP) [9] as a prior distribution. Because of its flexibility against large-scale, complex data, this framework is useful for cluster analysis and has been applied to a wide range of research fields such as natural language processing, image processing, and bioinformatics. As well as cluster analysis, topic analysis on grouped data, e.g., topic modeling with corpus data, has long been studied. The hierarchical Dirichlet process (HDP) [22] is an example of successful nonparametric Bayesian model for topic analysis. Used as a prior distribution of a mixture model, HDP extracts the mixture components (= topics) across groups and allows all topics to be shared by all groups, with mixture weights of topics inferred independently for each group. The following model discussion is based on document analysis. As such, words, documents, and topic, which are the expressions in document analysis, correspond to observations, groups, and mixture

components, which are generic technical expressions, respectively. The following model discussion can be applied to various fields (e.g., urban dynamics analysis [17]) in addition to the research fields mentioned above.

These two fields of study have developed independently, but considering that the cluster structure, or relationship among groups, enhances the performance of topic modeling described in [20], it is useful to treat these two analyses at the same time. The naive approach is to follow a sequential process. For example, first we extract topics using HDP and then cluster the document, or we cluster documents on the basis of tf-idf [12] and then extract topics for each document cluster. However, as shown in [24], the sequential process possibly suffers from inaccurate results because the optimization criteria of topic extraction and group clustering are different. Therefore, a nonparametric Bayesian model that simultaneously optimizes the topic extraction and group clustering as a unified framework is required.

As an alternative to such naive approaches, the nested Dirichlet process (nDP) [21] has been proposed. The nDP simultaneously extracts topics and clusters groups as a unified framework. In this model, groups (documents) of data are clustered into various clusters and topics are extracted for each cluster. Since the topics are not shared with groups in different clusters, there is a risk of over-fitting in the clusters to which few groups belong due to the lack of training data for the mixture components of such a cluster.

In order to solve this problem in nDP, Ma et al. [15] proposed a hybrid nested/hierarchical Dirichlet process (hNHDP). The hNHDP extracts global topics, which are shared by all clusters, and local topics, which are shared only by groups in the same cluster. Using the idea of [16], hNHDP clusters groups and allows partial topics (global topics) to be shared by all clusters. However, as with the nDP, this framework has the risk of over-fitting with regard to the cluster specific local topics of a cluster to which few groups belong due to the lack of training data for each topic. As mentioned in [15], enhancing the computational efficiency is also important, since the sampling method is used to infer the model parameters of hNHDP.

In light of this background, in this paper, we propose a coupled hierarchical Dirichlet process (cHDP) that archives the desired framework mentioned above in order to solve the problems that hNHDP is currently facing. The cHDP extracts topics and clusters groups as well as nDP and hNHDP and allows all mixture components to be shared by all clusters, as with HDP. In addition, in order to enhance the computational efficiency for handling large-scale data, we formulate cHDP so that it can use a variational Bayesian method in which analytical approximation is provided and convergence speed is improved compared to conventional sampling methods.

To evaluate our cHDP performance against the existing models, we conduct experiments with corpus data on topic modeling and document clustering. In addition, using large-scale mobility logs from smartphones, we apply the cHDP to big data analysis – in this case, urban dynamics analysis – in order to show that cHDP works well in the fields other than document modeling where the data

take continuous values, in contrast to the corpus data represented by discrete values. We perform experiments in which two simultaneous analyses are tackled: the extraction of the pattern of a daily transition of population common in target regions [17] and the clustering of these regions [25]. These analyses correspond to topic analysis and group clustering, respectively. As well as document modeling, since these two analyses have developed independently, and because even recent research [25] has proposed a sequential approach to such analysis, it is assumed that cHDP is useful in this urban dynamics analysis.

In order to clarify the position of our proposed cHDP, we introduce two existing models, nested hierarchical Dirichlet process (nHDP) [18] and coupled Dirichlet process (cDP) [13], whose names or motivation are similar to cHDP, and describe the differences between them and cHDP. The nHDP was proposed to extract tree structured, hierarchical topics, so unlike cHDP, it cannot realize simultaneous topic extraction or group clustering. In the case of cDP, its generic formulation is motivated by the same purpose as cHDP, but no concrete inference process was proposed in [13]. In this paper, we formulate a specific model equivalent to cDP and propose a closed-form variational inference that is superior to one in [13].

Our contributions are as follows. We developed a new nonparametric Bayesian method that simultaneously extracts topics and clusters groups in a unified framework while allowing all topics to be shared by all clusters. This is achieved by stochastic cluster assignment for both clustering processes. In order to enhance the computational efficiency, we formulate our model so that it can use a closed-form variational Bayesian method to approximately calculate the posterior distribution. We apply our proposed model to document analysis and big data analysis, in this case, urban dynamics analysis. The results of experiments with real data show that our model performs better in both research fields compared with existing models.

2 Related Works

As discussed in Sec.1, for grouped data, we propose a new framework that simultaneously extracts topics and clusters groups, which allows all mixture components (topics) to be shared by all clusters. In this section, we briefly describe the existing nonparametric Bayesian models for grouped data. First, we describe HDP as a basic model for grouped data that focuses on topic analysis and then we introduce nDP and hNHDP, which simultaneously do two analyses, as a baseline for comparison with our model. In the following explanation, we assume that we have D groups of data, and the nth observation of group d is denoted as $x_{d,n}$.

2.1 Model for Topic Analysis

HDP The hierarchical Dirichlet process (HDP) [22] is a nonparametric Bayesian model for grouped data. The generative process for a mixture model for grouped data is written as

$$G_0^* \sim \mathrm{DP}(\beta, H), \quad G_d \sim \mathrm{DP}(\alpha, G_0^*), \tag{1}$$

where $G_0^* \sim \mathrm{DP}(\beta, H)$ denotes the Dirichlet process (DP) [8], which draws discrete distribution G_0^*. β is a concentration parameter and H is a base measure of DP. This process is described by stick-breaking representation as

$$G_0^* = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \quad \phi_k \sim H, \quad \pi_k \sim \mathrm{GEM}(\beta), \tag{2}$$

where $\delta_\cdot$ is the Dirac's delta function. The expression GEM (named after Griffiths, Engen, and McCloskey [19]) is used as $\{\pi\}_{k=1}^{\infty} \sim \mathrm{GEM}(\beta)$ if we have $\pi_k = \pi'_k \prod_{j=1}^{k-1}(1-\pi'_j)$, $\pi'_k \sim \mathrm{Beta}(1, \beta)$ for $k = 1, \cdots, \infty$.

The group specific distribution G_d is drawn independently from $\mathrm{DP}(\alpha, G_0^*)$ and G_0^* is shared by all groups, which is itself drawn from another DP. As a result, mixture components (topics) are shared by all groups while the weights are independent of each group. The HDP cannot consider the relationship between groups, and since the mixture weights of each group are inferred independently, there is a risk of over-fitting.

2.2 Models that Simultaneously Extract Topics and Cluster Groups

NDP The nested Dirichlet process (nDP) [21] clusters groups and extracts topics in a unified framework. The nDP is written as the following process, in which the DP itself is used as the base measure of different DP:

$$Q \sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, H)), G_d \sim Q. \tag{3}$$

This generative process induces the clustering of groups. The mixture components and weights are shared only in the same cluster of groups. The stick-breaking representation of the nDP is written as

$$Q = \sum_{g=1}^{\infty} \eta_g \delta_{G_g^*}, \quad G_d \sim Q, \quad \eta_g \sim \mathrm{GEM}(\alpha), \tag{4}$$

$$G_g^* = \sum_{t}^{\infty} \pi_{g,t} \delta_{g,t}, \quad \phi_{g,t} \sim H, \quad \pi_{g,t} \sim \mathrm{GEM}(\beta). \tag{5}$$

Let G_g^* denote the cluster specific distribution and $\phi_{g,t}$ denote the tth parameter of cluster g. In the mixture model with the nDP, as the mixture components in a cluster are not shared by different clusters, the clusters to which few groups belong suffer from over-fitting due to the lack of training data.

HNHDP Ma et al. [15] proposed the hNHDP model, in which the advantages of the HDP and nDP are integrated. In the hNHDP, the cluster specific distribution F_g is modeled as the combination of two components, $G_0 \sim \mathrm{DP}(\alpha, H_0)$ and $G_g \sim \mathrm{DP}(\beta, H_1)$, and written as

$$F_g = \epsilon_g G_0 + (1 - \epsilon_g) G_g, \quad \epsilon_g \sim \mathrm{Beta}(\alpha, \beta). \tag{6}$$

G_0 is shared by all group clusters and G_g is cluster-specific. α, β are concentration parameters and H_0, H_1 are base measures. Therefore, we have global mixture components shared by all clusters and cluster-specific local mixture components. With this modeling, we can cluster the groups while some mixture components are shared by all clusters, which enhances the modeling performance. However, as well as the nDP, this framework still has the risk of over-fitting due to the cluster specific mixture components. To tackle this problem, we need a framework in which all mixture components are shared among all group clusters.

3 Coupled Hierarchical Dirichlet Process (cHDP)

As described in Sec.2, the existing nonparametric Bayesian models are facing various issues. In this section, we propose a coupled hierarchical Dirichlet process (cHDP) in which the advantages of HDP and nDP are integrated. The cHDP simultaneously extracts topics and clusters groups while allowing all mixture components to be shared by all group clusters, which solves the problem in the hNHDP. In addition, in order to enhance the computational efficiency, we modeled the cHDP so that it can use a variational Bayesian method in closed form for inferring the model parameters.

In this paper, we assume that we have D groups of data and let $\boldsymbol{x}_d = \{x_{d,1}, \ldots, x_{d,N_d}\}$ be the observations of group d, where $\{x_{d,n}\}$ denotes the nth observation and N_d is the total number of observations in group d. We assume that each observation $x_{d,n}$ is drawn from the probabilistic distribution $p(\theta_{d,n})$ with parameter $\theta_{d,n}$. The figure (D) in Fig.1 shows the generative process of cHDP.

3.1 Definition and Formulation

We define the generative process of our proposed cHDP as follows

$$G_0^* \sim \mathrm{DP}(\gamma, H),\ Q \sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, G_0^*)),\ G_d \sim Q. \tag{7}$$

The second equation of (7) indicates that the DP is used as the base measure of another DP as with the nDP described in (3). The base measure of the nested DP in (7) is drawn from another DP whose base measure G_0^* is shared with all groups as with HDP described in (1). Considering this description, we can say cHDP is the generative process that holds the characteristics of HDP and nDP.

Several representations such as the Chinese restaurant franchise and the stick-breaking process are candidates for implementing the cHDP. In this paper, we adopt the stick-breaking representation, which enables us to use variational Bayesian inference, a computationally efficient approximation method, because we consider using the cHDP to handle large-scale data. We formulate

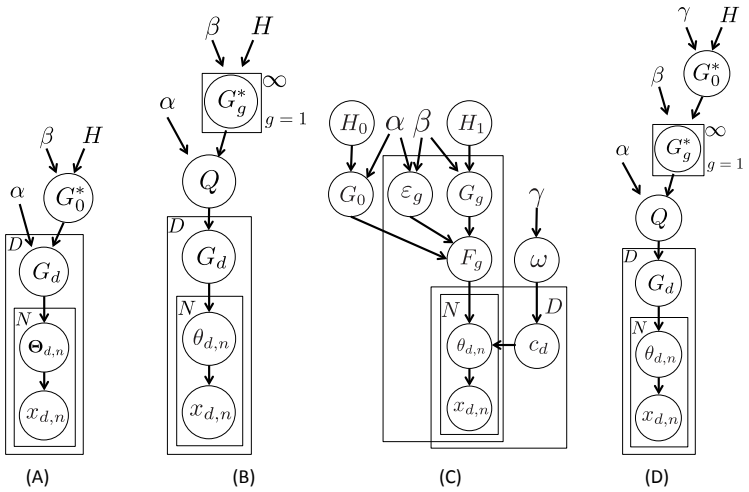

Fig. 1. Graphical model of (A) HDP, (B) nDP, (C) hNHDP, and (D) cHDP (proposed).

the stick-breaking representation of the cHDP as

$$G_0^* = \sum_{k=1}^{\infty} \lambda_k \delta_{\phi_k^*}, \quad \phi_k^* \sim H, \ \lambda_k \sim \mathrm{GEM}(\gamma), \tag{8}$$

$$G_g^* = \sum_{t}^{\infty} \pi_{g,t} \delta_{\psi_{g,t}^*}, \quad \psi_{g,t}^* \sim G_0^*, \ \pi_{g,t} \sim \mathrm{GEM}(\beta), \tag{9}$$

$$Q = \sum_{g=1}^{\infty} \eta_g \delta_{G_g^*}, \quad \eta_g \sim \mathrm{GEM}(\alpha), \quad G_d \sim Q, \tag{10}$$

where k is the index of mixture components shared by all groups and g is the index of the clusters of groups. Each group belongs to one of the clusters and cluster $g = 1 \cdots \infty$ has a cluster specific distribution G_g^* drawn as (9). Regarding the stick-breaking representation of the generative process of G_g^*, which is the same as the model structure of HDP in (7), there are different representations by Teh et al. [22] and Wang et al. [23]. The above representation is

$$G_g^* = \sum_{k=1}^{\infty} \pi_{g,k} \delta_{\phi_k}, \ \pi_{g,k} = \pi'_{g,k} \prod_{j=1}^{k-1} (1 - \pi'_{g,j}), \ \pi'_{g,k} \sim \mathrm{Beta}\left(\alpha \lambda_k, \alpha \left(1 - \sum_{j=1}^{k} \lambda_j \right) \right). \tag{11}$$

With this representation, it is not possible to use a variational method in closed form in the inference of posterior distribution, so we formulate as (9) using the representation in the same way as [23], which enables us to use the variational method. This is achieved by introducing cluster specific parameter $\{\psi_{g,t}\}_{t=1}^{\infty}$ and a mapping variable that connects $\psi_{g,t}$ and mixture component ϕ_k, which is shared by all clusters.

Next, we introduce additional variables and formulate the mixture model using the cHDP. Let $\mathbf{Y} = \{y_{d,g} | y_{d,g} = \{0,1\}, \sum_g y_{d,g} = 1\}$ be a variable that represents the cluster to which a group d belongs. Then, we define $\mathbf{Z} = \{z_{d,n,t} | z_{d,n,t} = \{0,1\}, \sum_t z_{d,n,t} = 1\}$ as a variable that represents the cluster specific component t to which $x_{d,n}$ belongs and $\mathbf{C} = \{c_{g,t,k} | c_{g,t,k} = \{0,1\}, \sum_k c_{g,t,k} = 1\}$ as a variable that represents the mixture component k to which the cluster specific component t of a cluster g corresponds. As mentioned above, introducing the cluster specific component t and mapping variable c enables us to use variational inference. Let $\boldsymbol{\Theta}$ denote the parameter set of distributions that the observations $\boldsymbol{X} = \{\boldsymbol{x}_{d,n}\}$ follow. The mixture model using the cHDP is then formulated as

$$p(\mathbf{X}|\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\Theta}) = \prod_{d,g,n,t,k} p(\boldsymbol{x}_{d,n}|\boldsymbol{\Theta}_k)^{y_{d,g} z_{d,n,t} c_{g,t,k}}, \tag{12}$$

$$p(\mathbf{Y}|\boldsymbol{\eta}') = \prod_{d,g} \left\{ \eta'_g \prod_{f=1}^{g-1} (1-\eta'_f) \right\}^{y_{d,g}}, \tag{13}$$

$$p(\mathbf{Z}|\mathbf{Y},\boldsymbol{\pi}') = \prod_{d,g,n,t} \left\{ \pi'_{g,t} \prod_{s=1}^{t-1} (1-\pi'_{g,s}) \right\}^{y_{d,g} z_{d,n,t}}, \tag{14}$$

$$p(\mathbf{C}|\boldsymbol{\lambda}') = \prod_{g,t,k} \left\{ \lambda'_k \prod_{j=1}^{k-1} (1-\lambda'_j) \right\}^{c_{g,t,k}}, \tag{15}$$

$$p(\eta'_g) = \mathrm{Beta}(\eta'_g|1,\alpha), \tag{16}$$

$$p(\pi'_{g,t}) = \mathrm{Beta}(\pi'_{g,t}|1,\beta), \tag{17}$$

$$p(\lambda'_k) = \mathrm{Beta}(\lambda'_k|1,\gamma). \tag{18}$$

3.2 Variational Bayesian Inference with Closed Form Update

As with the general nonparametric Bayesian models, the posterior distribution of this cHDP mixture model cannot be calculated in closed form. We therefore need to apply an approximation method such as Gibbs sampling or variational Bayesian inference. In this paper, because we consider application to large-scale data, we opt to use variational Bayesian inference, which is characterized by its computational efficiency, to approximately calculate the posterior distribution and infer the model parameters. We approximate the posterior distribution as

$$q(\cdot) \equiv q(\mathbf{Y})q(\mathbf{Z})q(\mathbf{C})q(\boldsymbol{\eta}')q(\boldsymbol{\pi}')q(\boldsymbol{\lambda}')q(\boldsymbol{\Theta}). \tag{19}$$

In variational inference, we update each parameter distribution q_i by $\ln q_i = \mathbb{E}_{q-i}[\ln p(\boldsymbol{X},\cdot)] + \mathrm{const.}$

Update $q(\mathbf{Y})$ We introduce $\xi_{d,g}$ that satisfies $\sum_g \xi_{d,g} = 1$ and

$$\ln \xi_{d,g} = \sum_{n,t} \mathbb{E}_q[z_{d,n,t}] \Bigg(\sum_k \mathbb{E}_q[c_{g,t,k}] \mathbb{E}_q[\ln p(\boldsymbol{x}_{d,n}|\boldsymbol{\Theta}_k)] + \mathbb{E}_q[\ln \pi_{g,t}] \Bigg) + \mathbb{E}_q[\ln \eta_g] + \text{const}, \tag{20}$$

then we have $q(\boldsymbol{y}_d) = \mathcal{M}(\boldsymbol{y}_d|\boldsymbol{\xi}_d)$ and $\mathbb{E}_q[y_{d,g}] = \xi_{d,g}$, where $\mathcal{M}(\cdot|\cdot)$ represents the multinomial distribution.

Update $q(\mathbf{Z}), q(\mathbf{C})$ As well as the update of $q(\mathbf{Y})$, both $q(\mathbf{Z})$ and $q(\mathbf{C})$ are represented as multinomial distribution by introducing variables.

Update $q(\boldsymbol{\eta}')$ We have $q(\eta'_g) = \text{Beta}(\eta'_g|\alpha_{g,1}, \alpha_{g,2})$, where

$$\alpha_{g,1} = 1 + \sum_d \mathbb{E}_q[y_{d,g}], \tag{21}$$

$$\alpha_{g,2} = \alpha_0 + \sum_{f=g+1}^{G} \sum_d \mathbb{E}_q[y_{d,f}]. \tag{22}$$

G is a large truncation number for group clusters. We also have

$$\mathbb{E}_q[\ln \eta'_g] = \psi(\alpha_{g,1}) - \psi(\alpha_{g,1} + \alpha_{g,2}), \tag{23}$$

$$\mathbb{E}_q[\ln (1 - \eta'_g)] = \psi(\alpha_{g,2}) - \psi(\alpha_{g,1} + \alpha_{g,2}), \tag{24}$$

$$\mathbb{E}_q[\ln \eta_g] = \mathbb{E}_q[\ln \eta'_g] \sum_{f=1}^{g-1} \mathbb{E}_q[\ln (1 - \eta'_f)], \tag{25}$$

where $\psi(\cdot)$ represents the digamma function $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$.

Update $q(\boldsymbol{\pi}'), q(\boldsymbol{\lambda}')$ As well as the update of $q(\boldsymbol{\eta}')$, both $q(\boldsymbol{\pi}')$ and $q(\boldsymbol{\lambda}')$ are represented as the beta distribution.

3.3 Predictive Distribution for New Observation

By using the approximation $p(\mathbf{C}, \boldsymbol{\eta}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{\Theta}|\mathbf{X}) \simeq q(\mathbf{C})q(\boldsymbol{\eta})q(\boldsymbol{\pi})q(\boldsymbol{\lambda})q(\boldsymbol{\Theta})$ as with [23], the likelihood of new observation $\boldsymbol{x}^*$ of the cHDP model trained with data $\boldsymbol{X}$ is written as

$$p^*(\boldsymbol{x}^*|\mathbf{X}) \simeq \sum_g \mathbb{E}_q[\eta_g] \prod_n \sum_t \mathbb{E}_q[\pi_{g,t}] \sum_k \phi_{g,t,k} \mathbb{E}_q[p(\boldsymbol{x}^*_n|\Theta_k)], \tag{26}$$

where

$$\mathbb{E}_q[\eta_g] = \mathbb{E}_q[\eta'_g] \prod_{f=1}^{g-1} (1 - \mathbb{E}_q[\eta'_f]),\ \mathbb{E}_q[\eta'_g] = \begin{cases} 1 & (g = G) \\ \frac{\alpha_{g,1}}{\alpha_{g,1} + \alpha_{g,2}} & (\text{o.w.}). \end{cases} \tag{27}$$

$\mathbb{E}_q[\pi_{g,t}]$ is calculated in the same manner.

4 Experimental Results

4.1 Document Analysis with Corpus Data

We present the experiments with corpus data to evaluate our framework. We constructed a topic model, cHDP-LDA, in which our cHDP is applied to latent Dirichlet allocation (LDA) [6] as a prior distribution. In the experiment with corpus, the words, documents, and topic correspond to observations, groups, and mixture components. The cHDP-LDA simultaneously optimizes both words and document clustering, and topics are shared by all document clusters.

Suppose we have document $d \in \{1, \cdots, D\}$ whose number of words is N_d and the total number of words found in these documents is W. Let $\boldsymbol{x}_{d,n} = \{x_{d,n,w} | x_{d,n,w} = \{0,1\}, \sum_w x_{d,n,w} = 1\}$ be the nth words in document d. We assume that the word $\boldsymbol{x}_{d,n}$ is drawn from multinomial distribution $\mathcal{M}(\boldsymbol{x}_{d,n}|\boldsymbol{\mu}_k)$, where k is the topic index and $\boldsymbol{\mu}_\cdot \in \mathbb{R}^W$ is a parameter of the multinomial distribution. The Dirichlet distribution $\mathcal{D}(\boldsymbol{\mu}|\boldsymbol{\delta}) \propto \prod_i \mu_i^{\delta_i - 1}$, which is conjugate to multinomial distribution, is used as a prior distribution for $\boldsymbol{\mu}$, where $\boldsymbol{\delta} \in \mathbb{R}^W$ is the hyperparameter for the Dirichlet distribution. In this paper, we assume that $\{\delta_i\}_{i=1}^W = \delta$ and $\mathcal{D}(\boldsymbol{\mu}|\boldsymbol{\delta})$ is the symmetric Dirichlet distribution.

In the following experiments, we used three corpora: Reuters-21578 Corpus (Reuters corpus) [10], Nist Topic Detection and Tracking Corpus (TDT2 corpus) [2], and NIPS Conference Papers Vols. 012 Corpus (NIPS corpus) [4]. With these datasets, preprocessing (removal of stop words, etc.) has already been done. For the Reuters corpus, we chose the version used in [3] composed of uniquely labeled documents with a total of 65 categories. The TDT2 corpus was collected from six news services from January 4, 1998 to June 30, 1998, and we chose the version used in [3] composed of uniquely labeled documents with a total of 96 categories. The NIPS corpus [4] was made with the proceedings of the Neural Information Processing Systems (Advances in NIPS) [1] from Vols. 0 (1978) to 12 (1999).

Perplexity Evaluation First, we evaluate the document modeling performance of our cHDP model and compare it to other existing topic models. All three corpora described above were used. As comparative models, we selected LDA models, each of whose prior distribution is an existing nonparametric Bayesian model, e.g., nested Chinese restaurant process (nCRP) [5], HDP [23], nDP [21], and hNHDP [15]. We refer to these models as hLDA, HDP-LDA, nDP-LDA, and hNHDP-LDA respectively. The hNHDP-LDA is a state-of-the-art framework that clusters both words and documents simultaneously. We set the hyperparameters of cHDP-LDA as $\alpha = \beta = \gamma = \delta = 1$, and those of nDP-LDA are also 1. As for hLDA, HDP-LDA and hNHDP-LDA, we followed the cited references.

We evaluate the models with the perplexity to test data. The perplexity indicates how well a trained model predicts new documents. Suppose we have D documents $\boldsymbol{X}^* = \{\boldsymbol{x}_d^*\}_{d=1}^D$ and the number of words in the dth document is N_d. In this case, the perplexity $\mathcal{P}(\boldsymbol{X}^*)$ is calculated as

$$\mathcal{P}(\boldsymbol{X}^*) = \exp\left(-\frac{\sum_d \ln p(\boldsymbol{x}_d^*)}{\sum_d N_d}\right). \tag{28}$$

The smaller the perplexity, the better the performance. In this experiment, we randomly divided each corpus into two groups, set A and set B, and then trained models with the one set and evaluated with the other.

For all corpora, the perplexities calculated with test sets A and B are shown in Table.1. The proposed cHDP-LDA performed best. The difference in performance between cHDP-LDA and HDP-LDA seems to be caused by the fact that cHDP can consider the relationship among documents. While the nCRP, which is the prior distribution of the hLDA, can indirectly consider the relationship of documents by partially sharing nodes (topics) in learning process, the hLDA performed worse than cHDP-LDA. We assume this is because the mixture weight to topics is independent of each document, resulting in over-fitting. HDP-LDA also suffers from this problem. Although the nDP-LDA can directly consider the relationship among documents, it exhibited a much worse performance than the others. This is because the topics in a document cluster to which few documents belong are inaccurate due to lack of training data, since topics in one document cluster are not shared by different clusters. The cHDP-LDA also outperformed the hNHDP-LDA, the state-of-the-art co-clustering model, in which partial topics are shared with different clusters. The same as the nDP-LDP, the hNHDP-LDA may suffer from over-fitting since hNHDP holds cluster specific topics (local topics). The above comparison clearly demonstrates that our cHDP-LDA, which clusters both words and documents while allowing all topics to be shared by all documents (or clusters), is suitable for topic modeling.

Table 1. Test data perplexity (best score in boldface).

Corpus	**Reuter**		**TDT2**		**NIPS**	
Training → Test	A → B	B → A	A → B	B → A	A → B	B → A
cHDP-LDA	**1591**	**1529**	**4157**	**4200**	**2543**	**2463**
hLDA	1925	1864	6523	5600	2584	2560
HDP-LDA	2478	2390	6348	6406	3033	2998
nDP-LDA	4557	4460	10043	10189	3404	3374
hNHDP-LDA	2041	1939	5498	5350	2886	2817

Document Clustering We conducted an experiment to evaluate only the performance of document clustering against the existing methods, some of which do not extract topics. The datasets used here are the Reuters corpus and the TDT2 corpus, both of whose documents are categorically labeled. The evaluation criterion is the adjusted Rand index (ARI) [11], which indicates the accuracy of the clustering result against the true labeling. If the clustering result coincides with the true labeling, ARI takes 1 and if the result is from random clustering, ARI takes 0. The closer the ARI value to 1, the better the clustering accuracy.

As comparative models, we used spherical k-means (SPK) [7] and spectral clustering (SC) [14], which cluster documents without topic extraction. In addition, as nonparametric Bayesian models, we used nDP and hNHDP. For each model, we conducted 100 clustering trials and evaluated the ARI values. Fig.2 indicates the means and standard deviation of ARI at each number of document clusters and Table.2 shows the highest ARI value and the corresponding number of clusters. In the case of cHDP-LDA, nDP-LDA, and hNHDP-LDA, since the number of document clusters is not manually determined (inferred by model), we plotted the same value for each number of document clusters.

We firstly compare the cHDP-LDA with SPK and SC, which do not extract topics. For the Reuters corpus, the ARI value statistically exceeded that of SPK and SC at the most appropriate number of document clusters. Although the ARI of the cHDP-LDA was slightly lower than that of SPK with the TDT2 corpus, the difference was not statistically significant. Then, we argue the result against the nDP-LDA and the hNHDP-LDA, nonparametric Bayesian models that cluster documents with topic extraction. Against the nDP-LDA, the cHDP-LDA statistically outperformed with both corpora. In contrast, although the cHDP-LDA performed slightly worse than the hNHDP-LDA for the Reuters corpus, without statistically significant difference, it statistically outperformed for the TDT2 corpus. We found the cHDP is more robust against documents than the HNHDP-LDA. These results indicate that the document clustering performance of the cHDP-LDA is the same level or higher compared to the existing methods.

We summarize the results of both experiments. As for the perplexity evaluation for topic modeling, our cHDP-LDA outperformed all existing models with all corpora. Regarding the ARI evaluation for document clustering, although cHDP-LDA performed slightly worse than some combinations of model and corpus, no statistically significant difference was observed by t-testing. In other cases, cHDP-LDA performed best and the difference was statistically significant for each case. Therefore, we conclude our cHDP-LDA performs better and more stably than other models including the hNHDP-LDA, the state-of-the-art model.

Table 2. Results of document clustering.

Reuters	No. of clusters	ARI
cHDP-LDA	—	0.419 ± 0.045
nDP-LDA	—	0.195 ± 0.103
hNHDP-LDA	—	$\mathbf{0.424 \pm 0.050}$
SPK	5	0.391 ± 0.109
SC	4	0.385 ± 0.019

TDT2	No. of clusters	ARI
cHDP-LDA	—	0.640 ± 0.028
nDP-LDA	—	0.083 ± 0.042
hNHDP-LDA	—	0.520 ± 0.066
SPK	12	$\mathbf{0.646 \pm 0.065}$
SC	7	0.557 ± 0.008

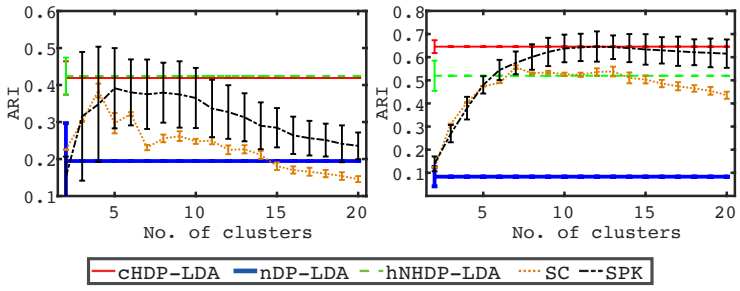

Fig. 2. Adjusted Rand indices with no. of clusters.

4.2 Big Data Analysis with Mobility Logs

In this section, using large-scale mobility logs from smartphones, we apply our cHDP to big data analysis, in this case, urban dynamics analysis. In this analysis, the following two analyses have been developed independently: extraction of patterns of the daily transition of population common in target regions [17], whose details are explained below, and clustering of regions [25]. Inspired by the success of cHDP in simultaneous topic modeling and document clustering, we apply cHDP to simultaneously tackle these analyses.

First, let us give an overview of this experiment. We set a square area (e.g., 300×300 m) as the target region and define this region as a point of interest (POI). In each POI, we divide a day into H time segments and describe the daily transition of population as a histogram, as shown in Fig.3. Each bin in the histogram is the number of logs observed in a time segment in the POI. We define basic patterns in the transition of population as dynamics patterns and assume that a daily transition of population is generated from the mixture of dynamics patterns. Using an analogy from document modeling, POI, a daily transition, and dynamics pattern correspond to document, word, and topic, respectively. Fig.4 shows the framework of this big data analysis by cHDP. The left side of the figure shows the collections of the daily transition of population in each POI and the right side indicates the extracted dynamics pattern.

Let d, n, and h be the index of POI, day, and time segment, respectively. The transition of population in the nth day in POI d is described as $\boldsymbol{x}_{d,n} = \{x_{d,n,1}, \cdots, x_{d,n,H}\} \in \mathbb{R}^H$. We assume $x_{d,n,h}$ is drawn from the mixture of Gaussian distribution and the distribution of the kth dynamics pattern is written as $\mathcal{N}(x_{d,n,h}|\mu_{k,h}, \rho_{k,h}^{-1})$. $\mu_{\cdot,\cdot}, \rho_{\cdot,\cdot}$ are the mean and precision. We use the Gaussian distribution and gamma distribution as the prior distribution for $\mu_{k,h}$ and $\rho_{k,h}$.

The dataset and the problem settings in this experiment are as below. We use the large-scale GPS logs collected from the disaster alert mobile application released by Yahoo! JAPAN. The logs are anonymized and include no users' information. Each record has three components: timestamp, latitude, and longitude. We use data collected for 365 days, from 1 July 2013 to 30 June 2014, consisting of 15 million logs per day in the Kanto region in Japan. We focus on

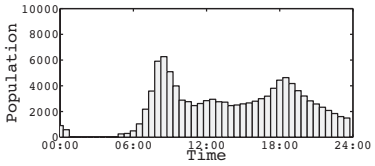

Fig. 3. Daily transition of population.

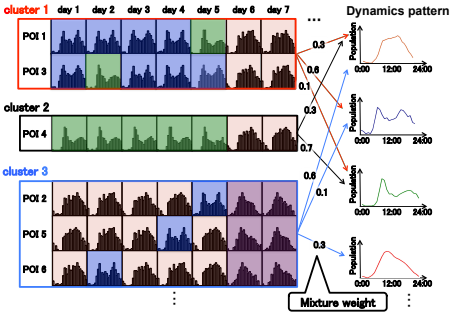

Fig. 4. Urban dynamics analysis by cHDP.

the square area (approximately 8000×8000 m) indicated by the thick blue line in Fig.6. We divide this focus area into 26×26 square pixels (each pixel is 300×300m) and regard each pixel as a POI. A daily transition of population in each POI is characterized by its scale and shape (e.g., the population peak time). As in [17], to make the patterns depend only on shape, we use the log counts divided by the average number of logs per day for training and test data for each POI.

For quantitative evaluation of dynamics pattern modeling, we use mean log likelihood (MLL) for test data. The models are trained with data of 30, 60, 90, 120, 150, and 180 days and tested by 180 days of data. From the 365 days of the dataset, training data and test data are randomly selected without duplication. Five tests are conducted with each number of days and the average values of MLL are evaluated. As for the evaluation for POI clustering, we visualize the clustering result and argue the validity on the basis of the real geographical features. This is because numerical evaluation is difficult for POI clustering.

We use the HDP and nDP as comparative models. Parameters are inferred by variational method. As for the POI clustering of HDP, we used a DP Gaussian mixture model with the mixture weight to dynamics pattern for each POI. Due to the computational performance for large-scale data, we do not use the hNHDP model, which is trained by sampling. Note that neither SPK nor SC can be directly used for region clustering without pattern extraction because feature value must be ratio scale calculated from the set of discrete values such as words.

Results As shown in Fig.5, the cHDP model had the best performance for all the training data condition. We can see a big performance gap between the cHDP and the others in the test with a small amount of training data. This result indicates that the cHDP's framework, i.e., considering the POI's relationship and the sharing dynamics patterns among all POIs, enhances the modeling accuracy. The reason nDP exhibited a worse performance is that the dynamics patterns in a POI cluster where few POIs belong are inaccurate due to the lack of training data, since patterns are not shared among different clusters.

Next, we evaluate the clustering performance. Since it is almost impossible to attach category labels by hand to such a small area, numerical evaluation like

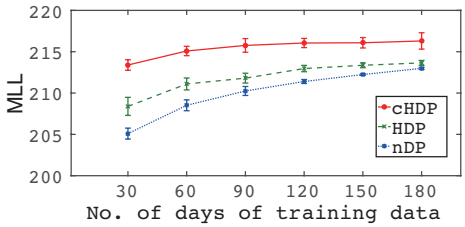

Fig. 5. Quantitative result of MLL.

ARI is difficult. Therefore, we visualize the clustering result and qualitatively argue the validity. Fig.7 shows the POI clustering result by the cHDP model. POIs that belong to the same cluster are drawn in the same color, while similar colors do not indicate the similarity in dynamics pattern trends. As shown in Fig.7, POIs distributed along railways are clustered into the same cluster (POIs around the Yamanote and Chuo lines are clustered in red and POIs around private railways are clustered in deep blue). In addition, yellow colored cluster corresponds to residential regions. Thus, it is shown that the cHDP model could cluster POIs corresponding to the actual geographical features.

The POI clustering by the HDP is shown in the left side of Fig.8. We first extracted dynamics patterns by the HDP and then clustered POIs on the basis of the mixture weights by DP. The correlation between the result and the actual geographical feature such as railways is low compared to the cHDP. In addition, neighboring POIs tended to belong to different clusters. Since we mesh the focus area into small areas (300×300 m), we assumed that spatial continuity of POI clusters among neighboring POIs can be seen. Therefore, the result is not valid and we cannot say that this is a meaningful clustering result. The comparison between cHDP and HDP indicates the advantage of simultaneous extraction of patterns and POI clusters. In contrast, as shown in the right side of Fig.8, the result of the nDP matches the geographical features to some extent. This is probably because the nDP simultaneously extracts patterns and clusters POIs as with cHDP. However, compared to the result of cHDP shown in Fig.7, POIs along the Yamanote and Chuo liens are not clustered well. We assumed that this difference stems from over-fitting of the cluster specific dynamics patterns. Considering the above evaluation, we conclude that the cHDP is useful for big data analysis, i.e., dynamics pattern extraction and region clustering.

5 Conclusion

In this paper, we proposed cHDP, a new nonparametric Bayesian mixture model that simultaneously extracts topics and clusters groups while allowing all topics to be shared by all clusters. In order to achieve better computational efficiency, we formulated our model in order to take variational Bayesian inference in closed form when inferring the model parameters.

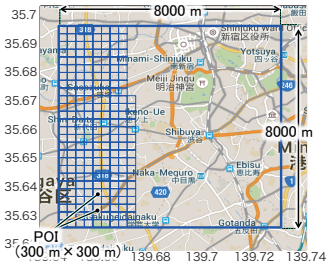

Fig. 6. Analysis area.

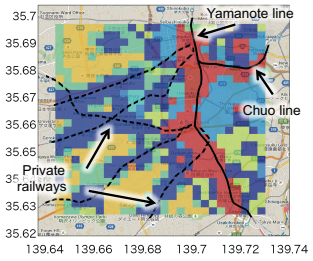

Fig. 7. Clustering by the cHDP model.

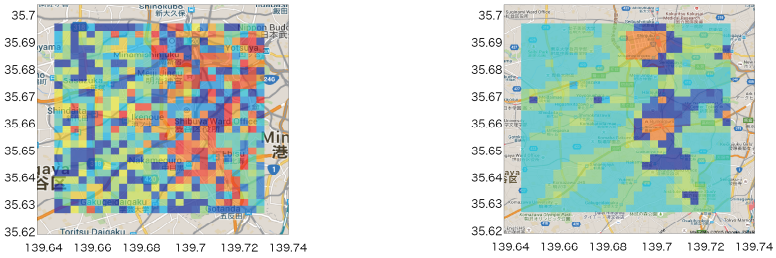

Fig. 8. Clustering by (left) HDP + DP clustering and (right) the nDP model.

We applied cHDP to document modeling and big data analysis, in this case, urban dynamics analysis. For the document modeling, we used cHDP as a prior distribution of LDA, which simultaneously conducts topic extraction and document clustering in a unified framework. Experiments with corpus data show that cHDP performs well in both tasks compared with existing models, achieving a 22% improvement against the state-of-the-art model. For big data analysis, we simultaneously tackled dynamics patterns extraction and region clustering. Using the GPS logs from smartphones, we showed that the cHDP enhances performance in pattern modeling and obtains valid clustering results. The comparison with nDP indicates the superiority of cHDP's topic sharing among all clusters.

For future work, we will introduce an online approach in the learning process. This is necessary to handle the data that accumulate over time, such as GPS logs from smartphones, let alone much more large-scale data. One option for this is using the online variational Bayesian method proposed in [23].

Acknowledgement. We thank Tengfei Ma, Issei Sato, and Hiroshi Nakagawa for providing the hNHDP implementation. This work was partly supported by CREST, JST.

References

1. Advances in Neural Information Processing Systems. http://books.nips.cc/ (visited 2013-01-15)

2. Nist Topic Detection and Tracking Corpus. http://projects.ldc.upenn.edu/TDT2/ (visited 2013-01-15)
3. Popular Text Data Sets in Matlab Format. http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html (visited 2013-01-15)
4. Sam Roweis: data. http://www.cs.nyu.edu/~roweis/data.html (visited 2013-01-15)
5. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM 57(2), 7:1–7:30 (2010)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine Learning research 3, 993–1022 (2003)
7. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Machine Learning 42(1-2), 143–175 (2001)
8. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230 (1973)
9. Ghahramani, Z., Griffiths, T.L.: Infinite latent feature models and the Indian buffet process. In: Proc. of NIPS. pp. 475–482 (2005)
10. Hayes, P.J., Weinstein, S.P.: CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In: Proc. of IAAI. pp. 49 – 64 (1991)
11. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2(1), 193–218 (1985)
12. Jones, K.S.: IDF term weighting and IR research lessons. Journal of Documentation 60(5), 521–523 (2004)
13. Lin, D., Fisher, J.: Coupled Dirichlet processes: Beyond HDP. In: Proc. of NIPS Workshop (2012)
14. Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (2007)
15. Ma, T., Sato, I., Nakagawa, H.: The hybrid nested/hierarchical Dirichlet process and its application to topic modeling with word differentiation. In: Proc. of AAAI. pp. 2835–2841 (2015)
16. Muller, P., Quintana, F., Rosner, G.: A method for combining inference across related nonparametric Bayesian models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(3), 735–749 (2004)
17. Nishi, K., Tsubouchi, K., Shimosaka, M.: Extracting land-use patterns using location data from smartphones. In: Proc. of Urb-IoT. pp. 38–43 (2014)
18. Paisley, J., Wang, C., Blei, D.M., Jordan, M.I.: Nested hierarchical Dirichlet processes. arXiv preprint arXiv:1210.6738 (2012)
19. Pitman, J.: Combinatorial stochastic processes. Tech. rep., Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course (2002)
20. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proc. of EMNLP. pp. 248–256 (2009)
21. Rodriguez, A., Dunson, D.B., Gelfand, A.E.: The nested Dirichlet process. Journal of the American Statistical Association 103(483), 1131–1154 (2008)
22. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)
23. Wang, C., Paisley, J.W., Blei, D.M.: Online variational inference for the hierarchical Dirichlet process. In: Proc. of AISTATS. pp. 752–760 (2011)
24. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception by hierarchical Bayesian model. In: Proc. of CVPR. pp. 1–8 (2007)
25. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proc. of KDD. pp. 186–194 (2012)