Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora

Jianwen Zhang\textsuperscript{1}, Yangqiu Song\textsuperscript{1}, Changshui Zhang\textsuperscript{1}, Shixia Liu\textsuperscript{2}
\textsuperscript{1}State Key Laboratory of Intelligent Technology and Systems
\textsuperscript{1}Tsinghua National Laboratory for Information Science and Technology
\textsuperscript{2}Department of Automation, Tsinghua University, Beijing 100084, China
\textsuperscript{2}IBM Research – China, Beijing, China
\{jw-zhang06@mails,zcs@mail\}.tsinghua.edu.cn; \{yqsong,liuxs\}@cn.ibm.com

ABSTRACT
Mining cluster evolution from multiple correlated time-varying text corpora is important in exploratory text analytics. In this paper, we propose an approach called evolutionary hierarchical Dirichlet processes (EvoHDP) to discover interesting cluster evolution patterns from such text data. We formulate the EvoHDP as a series of hierarchical Dirichlet processes (HDP) by adding time dependencies to the adjacent epochs, and propose a cascaded Gibbs sampling scheme to infer the model. This approach can discover different evolving patterns of clusters, including emergence, disappearance, evolution within a corpus and across different corpora. Experiments over synthetic and real-world multiple correlated time-varying data sets illustrate the effectiveness of EvoHDP on discovering cluster evolution patterns.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning; I.5.3 [Pattern Recognition]: Clustering; G.3 [Probability and Statistics]: Nonparametric statistics; H.2.8 [Database Management]: Database applications—Data mining

General Terms
Algorithms; Experimentation

Keywords
Multiple correlated time-varying corpora, clustering, mixture models, Bayesian nonparametric methods, Dirichlet processes

1. INTRODUCTION
Nowadays, we are surrounded by overwhelming quantities of textual materials from various heterogenous corpora (e.g., news, blogs) everyday. The themes in these corpora are usually similar, however, diversity also exists. For example, news typically has more discussions on society, politics, and economics than blogs which might focus more on personal life. Even within a corpus, the popularit of themes may also vary over time, and some of them may first appear in blogs, and then spread to news and message boards. Fig. 1 shows a real example of an evolving document cluster about “financial crisis” in three types of web corpora, including blogs, news and message boards.

To better understand the complex data, users not only want to examine the document clusters, but also want to discern the cluster evolution patterns over time and across corpora. Specifically, from multiple correlated time-varying corpora, it is desirable to discover the following patterns: (1) clusters within each corpus at each time epoch, (2) shared clusters among different corpora at each epoch, (3) evolving patterns of clusters within a corpus, and (4) evolving patterns of clusters across corpora over time.

However, it is challenging to learn cluster evolution patterns from such complex data. The first challenge is how to model the clusters both across different corpora and over time. On the one hand, we need a single integrated model for all corpora to set up a global bookkeeper of clusters, otherwise we can not easily discern the evolution of a cluster across corpora over time. On the other hand, different corpora may share some clusters while also having their distinctive clusters. Hence the commonality and diversity should be both reflected in the single integrated model. The second challenge is how to model the time dependencies in the multiple corpora setting. It is very common that the themes of a corpus evolves slowly along time, and thus the clustering patterns of adjacent time epochs usually exhibit strong correlations. Incorporating these correlations...
into the model in the multiple corpora setting is nontrivial. The last challenge is how to determine the cluster numbers. In time varying text data, a cluster may emerge and disappear. Consequently, the cluster numbers may change through time. It is awkward to require users to specify a cluster number at each time epoch for each corpus. Therefore a mechanism is preferred to automatically determine all the numbers of clusters.

The traditional clustering approaches deal with a single static data corpus. Hence the direct use of a general global clustering model on all data may fail to represent the diversity both over time and across different corpora. Beyond the classical clustering approaches, the most recent efforts often focus on tackling two sub problems. One is learning from multiple correlated text corpora, which aims to discover the related content across different text corpora as well as the distinctive information in each corpus [22, 26, 17]. Another is learning from a time-varying data corpus, which aims to discover the evolving patterns in the corpus as well as the snapshot clusters at each time epoch [3, 8, 9, 1, 16]. Both types of approaches are not sufficient to tackle the above challenges.

To deal with above challenges, in this paper, we propose an evolutionary hierarchical Dirichlet process (EvoHDP) model, which extends the hierarchical Dirichlet process (HDP) [22] to a time evolving scenario. In EvoHDP, each HDP is built for multiple corpora at each time epoch, and the time dependencies are incorporated into adjacent epochs under the Markovian assumption. Specifically, the dependency is formulated by mixing two distinct Dirichlet processes (DPs). One is the DP model for the previous epoch, and the other is an updating DP model. To infer the EvoHDP model, we also propose a casceded Gibbs sampling scheme. The proposed EvoHDP model can effectively discover cluster evolution patterns over time and across corpora. Moreover, the cluster numbers can automatically be determined due to the infinity property of DP.

2. RELATED WORK

In this section, we briefly introduce three categories of work related to this paper, including learning from multiple correlated data corpora, learning evolution patterns from a time-varying corpus, and some initial efforts involving multiple dynamic data.

In the research of learning from multiple correlated data corpora, HDP is a milestone [22]. It extended DP [2, 23] to model multiple correlated data corpora. In HDP, each data corpus is modeled by an infinite DP mixture model, and the infinite set of mixing clusters is shared among all corpora. Later works [16, 6] relax the assumption of HDP and incorporate more correlations between different corpora. Besides HDP, other efforts [4, 19, 26, 17] on this research topic are devoted to the extensions of the Latent Dirichlet Allocation (LDA) topic model [5]. However, none of them consider the problem of automatically determining the cluster/topic numbers. In fact, as pointed out in [22], HDP can be used as an LDA-based topic model, where the number of clusters can be automatically inferred from data. Therefore, HDP is more practical when users have little knowledge about the content to be analyzed.

In the research of learning evolutionary clusters from a time-varying corpus, evolutionary clustering [8, 9, 32, 1, 30, 31] is a new research topic. Evolutionary clustering aims to preserve the smoothness of clustering results over time, while fitting the data of each epoch. Among the above works, the approaches in [1, 30, 31] utilized DP to automatically determine the cluster numbers. In fact, incorporating time dependencies into DP mixture models is a hot topic in the research of Bayesian nonparametric [20, 10, 7, 16, 15, 14]. Moreover, some works have focused on extending LDA to dynamic topic models [3, 27, 25]. We noted that even though the title of [16] is similar to this paper, it actually presents an evolutionary DP mixture model for a single dynamic corpus.

To the best of our knowledge, there are four works that seemingly involve multiple time-varying data but actually handle different problems [28, 29, 11, 21]. Wang et al. [28] focused on detecting the simultaneous bustling of some topics in multiple text streams. They did not concentrate on the evolving patterns but only the bustling behavior of topics. Wang et al. [29] extended [28] to extract common topics from multiple text streams. They regarded that the underlying models of all streams are the same except that time delays exist between different streams. Hence they adjusted the timestamps of all documents to synchronize multiple streams and then learned a common topic model. Their assumption pays more attention to aligning topics of different streams and thus degenerates the topic diversity among different streams. Leskovec et al. [11] focused on tracking the spreading behaviors of short phrases (or “memes”) across the web to represent news cycles. Although memes act as signatures of clusters/topics, they are not enough to represent clusters/topics as the context information is lost. Tang et al. [21] worked on the dynamic multi-mode network with several types of actors. They aimed to provide a partition for each type of actors at each time epoch. There are two typical features in their problem: (1) the data to be handled is relational data, i.e., linkages between actors are required; (2) the partitions over time should be on a same set of actors while relationships among them vary over time. None of the above four works attempted to discover the cluster evolution over time and across corpora. On the contrary, this is exactly the focus of this paper. In addition, none of them handled the problem of automatically determining the cluster/topic numbers, which is one of the major considerations of our work.

3. PRELIMINARIES

In this section, we briefly introduce DP and HDP. A DP can be considered as a distribution of a random probability measure \( G \), and we write \( G \sim DP(\alpha_0, G_0) \), where \( \alpha_0 \) is a positive concentration parameter, and \( G_0 \) is a base measure. Sethuraman [18] showed that a measure drawn from a DP is discrete, by the following stick-breaking construction:

\[
\phi_1, \ldots, \phi_n \sim \text{Beta}(1, \alpha_0), \quad \pi \sim \text{GEM}(\alpha_0), \quad G = \sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}. 
\]

The discrete set of atoms \( \phi_1, \ldots, \phi_n \) are drawn from the base measure, and \( \text{GEM}(\alpha_0) \) refers to such a process: \( \pi \sim \text{Beta}(1, \alpha_0) \). We use boldface \( \pi = (\pi_1, \ldots, \pi_{\infty}) \) to represent the vector, which will be followed in the rest of the paper. \( \delta_{\phi_i} \) is a probability measure concentrated at \( \phi_i \). After observing draws \( \theta_1, \ldots, \theta_n \) from \( G \), the posterior of \( G \) is still a DP

\[
G|\theta_1, \ldots, \theta_n \sim DP \left( \frac{\theta_0 + n - 1}{\theta_0 + n}, \frac{m_0 \delta_\theta_0 + \alpha_0 G_0}{\theta_0 + n} \right).
\]

where \( m_0 \) is the number of draws in \( [\theta_1, \ldots, \theta_n] \), taking the same value \( \phi_\theta \). This posterior preserves the possibility of drawing a new distinct value from \( G_0 \) but puts more concentration on observed values.

HDP uses multiple DPs to model multiple correlated corpora. In HDP, a global measure \( G_0 \) is drawn from a DP \( (\gamma, H) \), with concentration parameter \( \gamma \) and base measure \( H \). Then, a set of measures \( G_j \) is drawn from a DP with base measure \( G_0 \). Each \( G_j \) models the corpus \( j \). Such a process is summarized as

\[
G_0 \sim DP (\gamma, H), \quad G_j | G_0, \alpha_0 \sim \text{DP}(\alpha_0, G_0).
\]

Given the global measure \( G_0, G_j \)'s are conditionally dependent. Having \( G_j, \pi_j \) data samples \( \{x_{i,j}\}_{i=1}^{n_j} \) in each corpus \( j \) are drawn from

1In general, we can regard the measure as a distribution.
We call the density parameterized by a distinct atom a “component”. This is the stick-breaking construction of HDP, and the corresponding graphical model is shown in Fig. 2(b).

4. EVOLUTIONARY HDP

In this section, we begin with the introduction of the EvoHDP model, and then show how to infer the model using a proposed Gibbs sampling technique.

We first introduce the data settings and some notations which are useful for subsequent discussions. There are $J$ corpora varying over $T$ time epochs. Considering the possibility that the data observed for some corpora at an epoch, we denote the number of corpora at epoch $t$ is $J_t$. At each epoch $t$, there are $n_f$ data samples in corpus $j$, and we denote a data sample (e.g., a document) as $x_{ji}$. We assume the underlying model to generate $x_{ji}$ for corpus $j$ at epoch $t$ is an infinite mixture model

$$p_j(x_j^t) = \int G_j^t(\theta) f(\theta|x_j) d\theta,$$

where $G_j^t = \sum_{i=1}^{\infty} \pi_{ij} \delta_{\phi_{ij}}$, and $f$ is the density of a distribution $F(\theta|x_j)$. We call the density parameterized by a distinct atom $\phi_{ij}$ as a mixing component, which describes a cluster.\footnote{We just call “hierarchical Dirichlet process mixture model” as HDP for short in this paper.}

4.1 Model

We model the multiple correlated time-varying corpora as a series of HDPs with time dependencies, as shown in the graphical representation of Fig. 3(a). Specifically, at each time epoch $t$, we use an HDP to model the multiple correlated corpora at that epoch and then put time dependencies between adjacent epochs based on the Markovian assumption. To build an overall bookkeeping of components for all epochs, we let these HDPs share an identical discrete base measure $G_0$, and $G$ is drawn from $\text{DP}(\xi, H)$ with $H$ as the base measure. We call $G$ the overall measure. Moreover, for an HDP to model the corpora at epoch $t$, we use $G_0$ to denote the global measure at that epoch, and call it the snapshot global measure. Then, the local measure $G_j^t$ for the $j$-th corpus at time epoch $t$ is called the snapshot local measure. In this way, an EvoHDP has one more layer than the original HDP [22].

The key issue of EvoHDP is how to incorporate time dependencies between adjacent epochs. We introduce two types of dependencies into different layers in EvoHDP to model the different evolving manner. In the following we will interpret this model step-by-step to explain why we use this scheme and how it would be useful.

The first is the dependency of snapshot global measure $G_0^t$ on $G_0^{t-1}$. Since $G_0$ is the measure for the components in all corpora at time $t$, the difference between $G_0^t$ and $G_0^{t-1}$ reflects the evolving of the global components in all corpora. We call this global time dependency.

The second is the time dependency within a corpus, i.e., the dependency of the snapshot local measure $G_j^t$ on $G_j^{t-1}$. Since $G_j^0$ is the measure for the components within corpus $j$ at time epoch $t$, the difference between $G_j^t$ and $G_j^{t-1}$ reflects the evolving of components within the corpus $j$. Then we call this intra-corpus time dependency.

In Fig. 3(a), we use dashed lines to represent the second type of dependencies, since in some cases (e.g., in HDP based LDA), there are no intra-corpus dependencies.

The generation process of EvoHDP is as follows.

1. Draw an overall measure $G \sim \text{DP}(\xi, H)$. $G$ plays a role of the overall component bookkeeping for all corpora at all epochs.

2. For each epoch $t$:

   2.1. Draw the snapshot global measure $G_j^t$ according to the overall measure $G$ and the previous snapshot global measure $G_0^{t-1}$:

   $$G_0^t \sim \text{DP}(\gamma', w'G_0^{t-1} + (1 - w')G_0^t),$$

   \hspace{1cm} (6)

   2.2. Draw the snapshot local measures $\{G_j^t\}_{j=1}^J$. Each $G_j^t$ for corpus $j$ at epoch $t$ is drawn according to the snapshot global measure $G_j^0$ and the previous snapshot local measure $G_j^{t-1}$:

   $$G_j^t \sim \text{DP}(\alpha_j^t, v_jG_j^{t-1} + (1 - v_j)G_j^0).$$

   \hspace{1cm} (7)

   2.3. For data samples $\{x_j^{t, w'}\}_{w'=1}^{v_j}$, draw the parameters of the component densities and generate the data samples:

   $$\theta_{ji}^{t, w'} \sim G_j^{t-1}, \quad x_{ji} \sim F(x|\theta_{ji}^{t, w'}),$$

   where $F(x|\theta_{ji}^{t, w'})$ is a distribution parameterized by $\theta_{ji}^{t, w'}$, e.g., that from an exponential family.

   Compared to the original HDP model, two levels of time dependencies are incorporated in by Eq. (6) and Eq. (7). When we set all $w'$ and $v_j$ to zero, the EvoHDP is a three-layer HDP.

Figure 2: Graphical representation for HDP: circles denote random variables, oval nodes denote parameters, shaded nodes denote observed variables, and plates indicate replication. (a) HDP. (b) The stick-breaking construction of HDP.

Figure 3: The graphical representation for the EvoHDP model. (a) The original representation. (b) The stick-breaking construction.
4.2 The Stick-Breaking Construction

According to the stick-breaking construction (Eq. (1)) of DP, we can write the explicit form of $G$:

$$G = \sum_{j=1}^{\infty} \nu_j \delta_{\theta_j}, \quad \nu \sim \text{GEM}(\xi).$$

Consequently, according to Eqs. (6) and (5), $G_0$ has the form

$$G_0 = \sum_{j=1}^{\infty} \beta_j^{(0)} \delta_{\theta_j}, \quad \beta_j^{(0)} \sim \text{DP}(\gamma^{(0)}, \beta_j),$$

where $\beta_j^{(0)} = \beta_j = \sum_{j=1}^{\infty} \beta_j^{(0)} \delta_{\theta_j}$, $\nu_j \sim \text{GEM}(\xi)$.

Similarly, we can also write the form of $G_j$ as

$$G_j = \sum_{j=1}^{\infty} \pi_j^{(j)} \delta_{\theta_j}, \quad \pi_j^{(j)} \sim \text{DP}(\nu_j^{(j)}, \pi_j^{(j)}),$$

where $\pi_j^{(j)} = \sum_{j=1}^{\infty} \pi_j^{(j)} \delta_{\theta_j}$.

In this way, we obtain the stick-breaking construction for EvoHDP, which is shown in Fig. 3(b), where $\pi_j^{(i)}$ is the index of the component emitting $x_j^{(i)}$.

According to this perspective, the EvoHDP provides a prior in which the snapshot models $[G_0]$, $[G_j]$, of all corpora at all epochs share the same infinite set of mixing components $\{\theta_j\}_{j=1}^{\infty}$. The differences among these snapshot models lie in the mixing weights.

4.3 Hierarchical Infinite Mixture Model and A Restaurant-Metaphor

Based on the stick-breaking construction for a DP, if we continue to use the stick-breaking constructions to represent $\pi_j^{(i)}$ and $\beta_j$ drawn from the two DPs in Eqs. (8) and (9), we obtain the hierarchical infinite mixture model of EvoHDP. This perspective clearly interprets the generation mechanism of EvoHDP.

We begin with the metaphor following the Chinese restaurant franchise (CRF) for HDP [22]. A corpus $j$ is called a restaurant, and a global atom $k$ is called a dish. We use a day to refer to a time epoch. We focus on the generation of component indicator $z_j^{(i)}$. Having $z_j^{(i)}$, the left generation process of $x_j^{(i)}$ is straightforward, which is drawn from $F(x|\theta_{z_j^{(i)}})$.

4.3.1 Generation of Snapshot Local Measure $G_j^{(i)}$

We first show the generation mechanism of each snapshot local measure $G_j^{(i)} = \sum_{k=1}^{\infty} \pi_j^{(i)} \delta_{\theta_j}$, i.e., the behavior of restaurant $j$ in day $t+1$. Since $\pi_j^{(i)}$ is drawn from $\text{DP}(\alpha_j^{(i)}, \pi_j^{(i)})$ as shown in Eq. (9), we can represent this DP using the stick-breaking construction as

$$\pi_j^{(i)} = \sum_{k=1}^{\infty} \pi_j^{(i)} \delta_{\theta_j}, \quad \{[k_j^{(i)}]_m \sim \text{iid} \beta_j^{(i)}, \quad \pi_j^{(i)} \sim \text{GEM}(\alpha_j^{(i)}),$$

which is illustrated in Fig. 4(a). We call $\tau$ a table, then $u_j^{(i)}$ is a distribution on tables. We can explain Eq. (10) as a peculiar way of dish serving in a restaurant. First, waiters place infinite number of tables. On each table $\tau$, a dish $k_j^{(i)}$ is selected from the local dish-meal menu $\hat{\pi}_j^{(i)}$ by waiters. Then, when a customer $i$ enters in restaurant $j$, he selects a table $\tau_j^{(i)}$ from the local customer-dish menu $u_j^{(i)}$ (with probability $u_j^{(i)}$) and enjoys the dish $k_j^{(i)}$, on the table. Consequently, the component indicator $z_j^{(i)} = k_j^{(i)}$. Moreover, $\pi_j^{(i)}$ plays the role of the local customer-dish menu, which indicates the customers’ preference on dishes in day $t+1$.

We then explain how a dish is placed on a table by the waiter. From Eq. (10), we see that $[k_j^{(i)}]_m \sim \text{iid} \beta_j^{(i)}$, i.e., a dish is drawn from the local table-dish menu $\hat{\pi}_j^{(i)}$. Notice that $\hat{\pi}_j^{(i)} = v_j^{(i)} \pi_j^{(i)} + (1 - v_j^{(i)}) \beta_j^{(i)}$ is a mixture of two menus, where $\pi_j^{(i)}$ is the local customer-dish menu in this restaurant yesterday and $\beta_j^{(i)}$ is the global franchise menu of current day recommended by the franchise manager. Then for a table $\tau$, a waiter select a dish $k_j^{(i)}$ from yesterday’s local customer-dish menu $\pi_j^{(i)}$ with probability $v_j^{(i)}$ while from current day’s global franchise menu $\beta_j^{(i)}$ with probability $1 - v_j^{(i)}$. This means in the franchise, a restaurant designs its localized menu by considering both yesterday’s local taste and current day’s franchise menu.

Obviously, it is possible that multiple tables have a same dish. In restaurant $j$ in day $t+1$, we denote the number of tables with dish $k$ as $T_j^{(i)}$. Then the total number of tables in restaurant $j$ in day $t+1$ is denoted as $T_j^{(i)} = \sum_{\tau \in \tau_j^{(i)}} T_j^{(i)}$. Then for a dish $k$, among the $T_j^{(i)}$ tables, there are $T_j^{(i)} k_j^{(i)}$ tables whose dishes are selected from yesterday’s local customer-dish menu $\pi_j^{(i)}$ and $T_j^{(i)} k_j^{(i)}$ tables whose dishes are selected from current day’s global franchise menu $\beta_j^{(i)}$. Then, $\forall k$, we have $T_j^{(i)} = T_j^{(i)} k_j^{(i)} + T_j^{(i)} k_j^{(i)}$. This mechanism is shown in Fig. 4(b).

4.3.2 Generation of Snapshot Global Measure $G_j^{(i)}$

Then we show the generation mechanism of the snapshot global measure $G_j^{(i)} = \sum_{k=1}^{\infty} \beta_j^{(i)} \delta_{\theta_j}$, i.e., how the franchise manager recommends the global franchise menu $\beta_j^{(i)}$ to all restaurants. The procedure is explained in Fig. 5(a). Since $\beta_j^{(i)}$ is drawn from $\text{DP}(\gamma_j^{(i)}, \beta_j^{(i)})$ as shown in Eq. (8), $\beta_j^{(i)}$ can also be represented using the stick-breaking construction as

$$\beta_j^{(i)} = \sum_{m=1}^{\infty} u_j^{(i)} \delta_{\theta_j}, \quad \{[k_j^{(i)}]_m \sim \text{iid} \beta_j^{(i)}, u_j^{(i)} \sim \text{GEM}(\nu_j^{(i)}).$$

We call each $m$ a metatable, then $u_j^{(i)}$ is a distribution on metatables. The manager has infinite number of empty metatables beforehand and he selects a dish $k_j^{(i)}$ for each metatable $m$ from the metatable-dish menu $\beta_j^{(i)}$ of day $t+1$. Remind that when a waiter in restaurant $j$ place a table $\tau$, with probability $1 - v_j^{(i)}$, he selects the dish $k_j^{(i)}$ according to the global franchise menu $\beta_j^{(i)}$. Now he just
need to select a metatable \( m^p_{t+1} \) according to the \( \text{waiter-metatable menu} \ u^{t+1} \), and then the dish \( k^p_{t+1} \) on the metatable is the one he should select. Hence the global franchise menu \( \beta^{t+1} \) also plays the role of the \( \text{waiter-dish menu} \), which indicates how a waiter selects dishes for tables.

The dishes on the metatables of the franchise in day \( t+1 \) also come from two menus, i.e., the yesterday’s global franchise menu \( \beta' \) and an \( \text{overall menu} \ \nu \). The overall menu \( \nu \) reflects common taste of the franchise. It is also possible that different metatables have a same dish and we denote the number of metatables with dish \( k \) as \( M_k^{t+1} \). Among the \( M_k^{t+1} \) metatables, there are \( M_k^{t+1} \) metatables whose dishes are selected from yesterdays’ franchise menu, and \( M_k^{t+1} \) metatables whose dishes are selected from the overall menu \( \nu \). Then \( \forall k \), we have \( \sum_{k=1}^{K} M_k^{t+1} = \sum_{k=1}^{K} M_k^{t+1} + \sum_{k=1}^{K} M_k^{t+1} \).

### 4.4 A Cascaded Gibbs Sampler

The hierarchical infinite mixture model and the restaurant franchise metaphor actually define a Gibbs sampling scheme for Evo-HDP. We can derive a cascaded Gibbs sampling procedure by sequentially sampling following variables.

#### 4.4.1 Sampling \( \nu \)

According to the generation of metatables introduced in Eq. 5(b), what are drawn from the overall measure \( G = \sum_{k=1}^{K} v_k \delta_{\beta_k} \) are the dishes on the metatables of all days designed by the franchise manager. We denote the number of all these metatables with dish \( k \) as \( M_k^t \), and the total number of metatables drawn from \( G \) as \( \sum_{k=1}^{K} M_k^t \). As \( G \sim \text{DP}(\xi, H) \), assume we have known the count variables \( \{M_k^t\} \), according to the property Eq. (2), the posterior of \( G \) is also a DP:

\[
G|\xi, H, \{M_k^t\} \sim \text{DP} \left( \xi + \sum_{k=1}^{K} M_k^t, \frac{H + \sum_{k=1}^{K} M_k^t \delta_k}{\xi + \sum_{k=1}^{K} M_k^t} \right)
\]

where \( K \) is the number of distinct dishes on all metatables. According to Sec. 5.2 of [22], \( G \) can be represented as

\[
G = \sum_{k=1}^{K} v_k \delta_{\beta_k} + v_u G_u, \quad G_u \sim \text{DP}(\xi, H) \tag{12}
\]

\[
\nu = (v_1, \ldots, v_K, v_u) \sim \text{Dirichlet}(M_1, \ldots, M_K, \xi). \tag{13}
\]

This augmented representation reformulates original infinite vector \( \nu \) to an equivalent finite one with length \( K + 1 \). Then \( \nu \) is sampled from the Dirichlet distribution of Eq. (13).

Notice that in the following, \( G_0 ^t \) and \( G_1 ^t \) are also represented using above augmented representation as

\[
G_0 ^t = \sum_{k=1}^{K} \beta_k \delta_{\beta_k} + \sum_{k=1}^{K} \beta_k^t \delta_{\beta_k^t}, \quad G_1 ^t = \sum_{k=1}^{K} \pi_k \delta_{\beta_k} + \pi_k^t \delta_{\beta_k^t}, \quad G_u \sim \text{DP}(\xi, H).
\]

Then \( \beta \) and \( \pi^t \) are both represented as finite vectors

\[
\beta = (\beta_1, \beta_2, \ldots, \beta_K), \quad \pi^t = (\pi^t_1, \ldots, \pi^t_K, \pi^t_u).
\]

#### 4.4.2 Sampling \( \beta \) and \( \pi^t \)

According to Fig. 5(b), what are drawn from \( \beta' \) include two parts. One part is the \( T_{\beta'}^{t+1} = \sum_{k=1}^{K} T_{\beta_k^t} \) dishes on the \( T_{\beta'}^{t+1} \) tables in all restaurants of day \( t \). The second part is the \( M_{\beta'}^{t+1} \) metatables of next day \( t + 1 \). We call all these tables and metatables drawn from \( \beta' \) as pseudo-tables. We denote the number of pseudo-tables with the same dish \( k \) as \( T_{\beta_k^t} \), then \( T_{\beta'}^{t+1} = \sum_{k=1}^{K} T_{\beta_k^t} + M_{\beta'}^{t+1} \). As \( \beta' \sim \text{DP}(\gamma', \tilde{\beta}') \), and assuming we have obtained the count variables \( \{T_{\beta_k^t}\} \), the posterior of \( \beta' \) is also a DP. Similar to Eq. (13), \( \beta' \) can also be sampled from a Dirichlet distribution

\[
\beta' = (\beta_1', \beta_2', \ldots, \beta_K'), \quad \tilde{\beta}' = (\tilde{\beta}_1', \ldots, \tilde{\beta}_K').
\]

where \( \gamma' = \gamma + T_{\beta'}^{t+1} \), and

\[
\beta_1' = \frac{1}{\gamma'} \left( \gamma' w' \tilde{\beta}_1' + \gamma' (1 - w') \gamma_1 + T_{\beta_1'}^{t+1} \right),
\]

\[
\beta_k' = \frac{1}{\gamma'} \left( \gamma' w' \tilde{\beta}_k' + \gamma' (1 - w') \gamma_k + T_{\beta_k'}^{t+1} \right), \quad k = 2, \ldots, K.
\]

The sampling of \( \pi^t_k \) is similar to that of \( \beta' \):

\[
\pi^t_k \sim \text{Dir}(\tilde{\pi}_1^t, \ldots, \tilde{\pi}_K^t).
\]

#### 4.4.3 Sampling \( z_{jk} \)

Given \( \pi^t_k \), sampling \( z_{jk} \) is straightforward:

\[
p(z_{jk} = k | x_{jk}, \ldots, x_{jk}) = 1 \cdot \frac{1}{(1 - \alpha_k^t) + \alpha_k^t \sum_{\nu=1}^{\nu_u} \varphi_{\nu}},
\]

where \( x_{jk}^t \) is the set of all the samples having been assigned to component \( k \), other than \( x_{jk}^t \).
4.4.5 Sampling Hyper-parameters

The concentration parameters of DPs, i.e., $\xi$, $[\gamma^t]$ and $[\alpha^t_0]$, can also be sampled by putting a vague gamma prior on them

$$\xi \sim \text{Ga}(a_0, b_0), \gamma^t \sim \text{Ga}(a_0, b_0), \alpha^t_0 \sim \text{Ga}(a_0, b_0).$$  (28)

The sampling method is the same as that in [22]. Moreover, the time dependency parameters $\omega^t$ and $\nu^t$ can be taken as controlling parameters or also sampled using the method in [16] by putting a Beta prior for them and sampling from the posterior.

According to the sampling method for groups of variables described above, there are recursive dependencies along hierarchies and time epochs. We follow the dependencies of different sets of variables and design a cascaded Gibbs sample scheme. The procedure is summarized in Algorithm 1.

Algorithm 1 A cascaded Gibbs sampling scheme (one iteration)

1: for $t = T$ to 1 do
2:    for $j = 1$ to $J$ do
3:       Sample $Z^t_j$ according to Sec. 4.4.3.
4:       $\forall k = 1, \ldots, K$, sampling count variables $T^t_j → z^t_j, T^t_j → z^t_j$, and $T^t_j$ according to Eq. (22-24).
5:    end for
6:    $\forall k = 1, \ldots, K$, sampling count variables $M^t_j → z^t_j, M^t_j → z^t_j$, and $M^t_j$ according to Eq. (25-27).
7: end for
8:  Sampling concentration parameters $\xi, \gamma^t$ and $\alpha^t_0$.
9:  Sampling $\nu$ according to Sec. 4.4.1.
10: for $t = 1$ to $T$ do
11:     Sampling $\phi^t_j$ according to Sec. 4.4.2.
12:     for $j = 1$ to $J$ do
13:        Sampling $\phi^t_j$ according to Sec. 4.4.2.
14:    end for
15: end for

4.5 Global and Local Components

We not only need the component assignments $Z$, but also the component parameters $\{\phi^t_j\}$. As introduced in Eq. (21), $\{\phi^t_j\}$ have been integrated out in the sampling process. Having assignments $Z$, we can obtain the posterior of a $\phi^t_j$:

$$p(\phi^t_j | x^t_j, z^t_j = k, \forall \ell, j, i), H) \propto p(\phi^t_j | H) p(\{x^t_j, | z^t_j = k\} | \phi^t_j).$$

This distribution is a “global” one conditioned on data of all corpora from all epochs. In textual data, it denotes a global component $k$ in the entire textual collection. If we limit the data in a corpus $j$ at epoch $t$, we also obtain the posterior of $\phi^t_j$ as a local component

$$p(\phi^t_j | x^t_j, z^t_j = k, \forall i), H) \propto p(\phi^t_j | H) p(\{x^t_j, | z^t_j = k, \forall i\} | \phi^t_j).$$

This is the component $k$ as a local one in corpus $j$ at epoch $t$.

5. EXPERIMENTS ON SYNTHETIC DATA

In this section, we use a synthetic data set from mixtures of multinomial distributions to test our approach.

5.1 Data

The data set consists of three time-evolving corpora covering 5 time epochs. Hence $t, J$, the number of corpora at epoch $t$, is 3. There are totally $K = 8$ components (dishes) involved in all corpora. Each dish $k$ is a 2-dimensional multinomial distribution $F(x|\phi^t_j) = \text{Multinomial}(x; \phi^t_j)$, with density

$$f(x|\phi^t_j) = \frac{\prod_{d=1}^D x_d^t !}{\prod_{d=1}^D x_d !} \prod_{d=1}^D \phi^t_{j,d},$$

where $x$ is a $D$-dimensional nonnegative integer vector, and $\phi^t_j$ is a $D$-dimensional nonnegative real vector with $\sum_d \phi^t_{j,d} = 1$. Here $D = 2$ and we set $\sum_d x_d = 200$. All the $\phi^t_j$’s are listed in Tab. 1. Each corpus $j$ at epoch $t$ is a uniform mixture of 3 tables, and each table $\tau$ is associated with a dish indexed by $k^t_j$. We denote the dish associated with table $\tau$ as $k^t_j$. Hence this local model can be represented by

$$p^t_j(x) = \sum_{\tau = 1}^3 \frac{1}{3} \text{Multinomial}(x; \phi^t_{\tau,j}).$$

Then $\phi^t_j$ samples are drawn from this mixture model, which compose the corpus $j$ at epoch $t$. More details of the data set are shown in Tab. 1. In the conjunction area of “Tables ($k^t_{j1}, k^t_{j2}, k^t_{j3}$)”, the triple at row $t$ and column $j$ are the three dish indices to compose the local mixture model. In such a data set, different corpora overlap on some components, and the underlying models evolve over time.

5.2 Evaluation Criteria

We introduce several numerical criteria to evaluate two types of performances. The first type is the static performance on fitting training data and predicting held-out data. The second type is the temporal performance on preserving correlation between epochs, including the correlation within a corpus and that across different corpora.

5.2.1 Static Criteria

Two criteria are used to evaluate the static performance, the normalized mutual information (NMI) and perplexity.

NMI measures coherence between the clustering assignments and the true category labels. A higher value on NMI indicates a better clustering result. For each corpus $j$ at each epoch $t$, having component assignments for all data samples, we can compute the value $NMI$ for the corpus. Then we use average value $\sum_{j \in J} NMI_j / \sum_{j \in J} 1$ as the final result on the criterion of NMI.

Perplexity is a standard metric in information retrieval. We denote the training set as $X_{\text{train}}$ and the held-out test set as $X_{\text{test}}$, then the per-sample perplexity of a model is defined based on the likelihood of “generating the test set given the training set”:

$$\text{Perplexity} = \exp \left( -\frac{1}{n_{\text{test}}} \sum_{i,j} \log p(x^t_{i,j}|\text{Model}, X_{\text{train}}) \right),$$

where $x^t_{i,j}$ is the $i$-th data sample in corpus $j$ at time epoch $t$ and $n_{\text{test}}$ is the size of test data set. In text modeling, to eliminate the fluctuation caused by the different lengths of documents, the perword-perplexity is often used instead, i.e., $n_{\text{test}}$ is the number of all the tokens in the test document collection. In this paper, we use log(perword-perplexity) and call it LogPerp. A lower value on LogPerp indicates better prediction performance.

5.2.2 Temporal Criteria

We define three types of temporal criteria to evaluate the time dependency and model smoothness between time epochs.

- **Intra-corpus temporal correlation (IntraCorr)** is defined to describe the average correlation between adjacent epochs within a corpus:

$$\text{IntraCorr} \triangleq \sum_{j \in J} E \left[ (\phi^t_j)^\top (\phi^{t+1}_j) \right] / \sum_{j \in J} 1.$$

- **Inter-corpora temporal correlation (InterCorr)** is defined to describe the average correlation between different corpora of adjacent epochs:

$$\text{InterCorr} \triangleq \sum_{j \in J} E \left[ (\phi^t_j)^\top (\phi^{t+1}_j) \right] / \sum_{j \in J} 1, l \neq j.$$

- **Global temporal correlation (GCorr)** is defined to describe the correlation between global distributions $G_{\tau}$ of adjacent epochs:

$$\text{GCorr} \triangleq \sum_{\tau} E \left[ (\phi^t)^\top (\phi^{t+1}) \right] / \sum_{\tau} 1.$$

Above three criteria measure the time dependencies from the as-
pect of correlation. Higher values on them are favored when static performances are similar. We also define another three types of criteria from the aspect of divergence between adjacent epochs’ distributions. On the contrary, lower values on them are favored when static performances are similar. They are defined as follows.

**Intra-corpus temporal KL divergence (IntraKL)**

$$\text{IntraKL} \doteq \sum_{i,j} \mathbb{E} \left[ KL\left( \phi_i^t \| \phi_j^{t+1} \right) \right] / \sum_{i,j} 1.$$  

**Inter-corpora temporal KL divergence (InterKL)**

$$\text{InterKL} \doteq \sum_{i,j} \mathbb{E} \left[ KL\left( \phi_i^t \| \phi_j^{t+1} \right) \right] / \sum_{i,j} 1, \ i \neq j.$$  

**Global temporal KL divergence (GKL)**

$$\text{GKL} \doteq \sum_{i} \mathbb{E} \left[ KL\left( \beta^i \| \beta_t^{t+1} \right) \right] / \sum_{i} 1.$$  

where $KL(\|)$ is the KL-divergence between two distributions.

In all the temporal criteria defined above, the expectations are calculated by MCMC samples obtained during the cascaded Gibbs sampling process.

### 5.3 Settings and Results

Except for EvoHDP, we also ran a three-layer HDP, which is just a special case of EvoHDP when all the time dependencies are removed. All the settings for EvoHDP and HDP are the same. The component model $F(t) \phi$ was set to a multinomial distribution, and the base measure $H$ was set to the conjugate prior for $F$, a symmetric Dirichlet distribution with parameter 0.5. The vague gamma priors in Eq. (28) for the concentration parameters of EvoHDP and HDP were set to be $Ga(10,0,1.0)$. An identical set of randomly generated component assignments was used to initialize both EvoHDP and HDP. In addition, to study the impact of time dependency parameters $\omega$ and $\nu$, we also set $\nu = \nu’ = \omega$ as a controlling parameter and swept $\omega$ in $[0.1, 0.3, 0.5, 0.7, 0.9]$. We call the EvoHDP model with $\omega = \nu’ = \omega$ as “EvoHDP”, and call the EvoHDP model with $\nu’$ also sampled during the inference as “EvoHDP-spl”.

For the Gibbs sampling procedure of both models, we ran 20 chains, and set the burn-in time as 1000 for each chain. After the burn-in, from each chain another 500 MCMC samples were preserved, then we obtained 1000 MCMC samples to calculate the evaluation criteria. The models were evaluated via 10-fold cross validation, and all criteria results were averaged on the 10 rounds.

### 6. EXPERIMENTS ON REAL DATA

In this section, we report the experiments on a real online document collection. This data set consists of 103,986 text articles queried from a search engine, Boardreader\(^3\), in which the time stamps of the articles are ranged from July 2008 to December 2008, using 20 financial companies’ names, e.g., “AIG insurance”, “Bank of America”, “State Farm”, etc. All these articles come from three types of public websites, i.e., news, blogs and message boards. We used the Mallar [13] package to pre-process the data set. We removed the stop words and rare words (appearing less than 10 times in the whole collection). After that, the vocabulary size of this text data set was $W = 77,999$. Term frequencies were extracted to represent each article. We organized the data set into

\[http://boardreader.com/\]
These five clusters are all finance related. For example, the “electric information”, “Barclays premier league” and “bank related”.

To have a clear insight into the evolution patterns in the multiple views. The three clusters “Barclays premier league”, “election”, and “financial crisis” are shown in Fig. 12(b), Fig. 12(a) and Fig. 1, respectively. From the three figures, besides clearer witness of the phenomenon of financial crisis in this month. More such features can also be observed in all the three figures. Second, the most frequent keywords of a cluster also vary over time, reflecting the evolving of the content of a cluster. Taking the cluster “election” (Fig. 12(a)) as an example, in blogs, in August, the keywords “Obama”, “Bush”, “presidential”, “democratic” indicated the normal features before an election. However, in September, “McCain” appeared as the hottest keywords as Republicans nominated John McCain for president in September 4th. Besides, crisis related keywords such as “crisis”, “aig”, “(wall) street” became hot due to the break-out of financial crisis in this month. More such features can also be observed in later months and other clusters. (5) *Cluster evolution across different corpora.* In Fig. 1 and Fig. 12(a), it is clear that both “crisis” and “election” clusters were first active in blogs, and then became popular in news and message boards.
Figure 11: Clusters in each corpus. For the two rows of figures, the meaning of the height of a colored stripe is different. In the first row, the width is the number of articles assigned to the cluster, i.e., $n_{jk}$, while in the second row, it is the proportion of the cluster at that epoch, i.e., $\pi_{jk}$. 
7. CONCLUSIONS

We propose an evolutionary hierarchical Dirichlet process (EvoHDP) model to mine cluster evolution from multiple correlated time-varying corpora. EvoHDP extends original HDP by incorporating time dependencies into a series of HDPs. A cascaded Gibbs sampling scheme is proposed to infer the model. Our approach can discover cluster emergence, disappearance, and evolution within a corpus and across different corpora. In addition, the cluster numbers for all corpora at all epochs are inferred from data rather than specified.

Experiments on a synthetic data set and a real-world financial related web data set validated the effectiveness of our approach. Compared to the original HDP, EvoHDP exhibits better predicting ability and stronger correlations across corpora over time on both data sets. In addition, on the real financial related web data, we observed that the cluster evolution patterns, emergence, disappearance, evolution within a corpus and across corpora, can be effectively discovered by EvoHDP.

8. ACKNOWLEDGEMENTS

The authors Jianwen Zhang and Changshui Zhang were supported by National Natural Science Foundation of China (NSFC, Grant No. 60835002). We would like to thank Weihong Qian and Furu Wei for their help on preparing the visualization results. We also thank all the reviewers for the suggestions to improve the paper.

9. REFERENCES


