Discovering Hierarchical Topic Evolution in Time-Stamped Documents

Jun Song
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, and University of Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: songjun210@mails.ucas.ac.cn

Yu Huang
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: huangyu23@sina.com

Xiang Qi
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, and University of Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: qixiang09@mails.ucas.ac.cn

Yuheng Li
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China, and University of Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: liyuheng2012@gmail.com

Feng Li
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: lifeng@ie.ac.cn

Kun Fu
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: kunfuiecas@gmail.com

Tinglei Huang
The Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, No. 9 Zhongguancun Beiyitiao Alley, Haidian District, Beijing 100190, China. E-mail: tlhuang@mail.ie.ac.cn

The objective of this paper is to propose a hierarchical topic evolution model (HTEM) that can organize time-varying topics in a hierarchy and discover their evolutions with multiple timescales. In the proposed HTEM, topics near the root of the hierarchy are more abstract and also evolve in the longer timescales than those near the leaves. To achieve this goal, the distance-dependent Chinese restaurant process (ddCRP) is extended to a new nested process that is able to simultaneously model the dependencies among data and the relationship between clusters. The HTEM is proposed based on the new

Received March 6, 2014; revised September 26, 2014; accepted September 26, 2014

 DOI: 10.1002/asi.23439

process for time-stamped documents, in which the time-stamp is utilized to measure the dependencies among documents. Moreover, an efficient Gibbs sampler is developed for the proposed HTEM. Our experimental results on two popular real-world data sets verify that the proposed HTEM can capture coherent topics and discover their hierarchical evolutions. It also outperforms the baseline model in terms of likelihood on held-out data.

Introduction

The explosive growth of electronic document collections has created great interest in automatic text analysis methods. Among these, the topic evolution method is an important one that focuses on discovering what and how topics change over time from time-stamped documents (Can et al., 2010; He et al., 2009; Wilkinson & Thelwall, 2012). The discovered evolving topics are able to concisely reflect the dynamic contents of the documents.

A number of topic models have been proposed in the recent past for topic evolution (Blei & Lafferty, 2006; Lin et al., 2012; Sugimoto, Li, Russell, Finlay, & Ding, 2011; Wang & McCallum, 2006). In the framework of topic models, each document is projected into a distribution over topics that capture the semantic contents of the documents (Blei, Ng, & Jordan, 2003). Topic models that incorporate time have become popular choices for topic evolution. Among these models, topic over time (TOT; Wang & McCallum, 2006) was the first to model the words and timestamps jointly in a probabilistic graphical model. In TOT, a beta distribution over time is associated with each topic to capture popularity. This modeling of timestamp helps to discover more explicit topics as well as their popularity, and reveal when and over what length of time the topical trends are occurring. Another model called the recurrent Chinese restaurant process (RCRP; Ahmed & Xing, 2010) takes a slightly different path to incorporate time. The RCRP discretizes document timestamps into epochs and assumes that the current popularity of a topic depends on its past popularity. Therefore, it captures topic changes in a more flexible manner. Note that the RCRP is capable of modeling the birth, growth, and death of topics. Apart from these two models, there are many other topic models proposed for topic evolution, such as the nonparametric TOT (Dubey, Hefny, Williamson, & Xing, 2013), the Storyline (Ahmed et al., 2011), and so on. All of these models have been of great help for document analysis and organization.

However, there are still some unsolved problems in topic evolution, one of which is that existing topic evolution models ignore the relationship between evolving topics. Previous literature has discovered that topics are closely related to each other (Blei, Griffiths, & Jordan, 2010; Blei & Lafferty, 2007). For example, a document about genetics is more likely to also be about disease than X-ray astronomy (Blei & Lafferty, 2007). This kind of relationship comes largely from the co-occurrence of words in the documents. Modeling this kind of relationship is useful for improving the coherence of topics, and it also provides an easy way to guide the user from one topic to other related topics. In other words, ignoring this kind of relationship might lead to disordered topics, especially for large document collections.

The objective of this paper is to model the relationship between evolving topics. To achieve this goal, we first extend the distance-dependent Chinese restaurant process (ddCRP; Blei & Frazier, 2011) to a new nested process named nddCRP. The ddCRP is able to cluster various dependent data with the prior that nearby data are more likely to be clustered together. The new process inherits this advantage, as well as providing a relationship between clusters by arranging clusters in a hierarchy. Then a hierarchical topic evolution model (HTEM) is proposed based on the new process. In the proposed HTEM, each node of the tree represents a topic and the difference of timestamps is utilized to measure the distances between documents. According to the tree structure, topics near the root of the tree are more abstract than those near the leaves. Moreover, various timescales can be set in the tree to discover multiscale topic evolution. The effectiveness of the HTEM is evaluated by two popular real-world data sets. One is the presidential State of the Union addresses over 200 years and the other is the Neural Information Processing Systems (NIPS) conference papers across 13 years. The experimental results verified that the proposed model is able to discover meaningful topic evolutions in a hierarchy and also to obtain a better likelihood than the baseline on held-out data.

The remainder of this paper is organized as follows. The next section discusses related works. The Hierarchical Topic Evolution Model is then proposed in more detail. The experimental results are then presented for the two popular real-world data sets. Finally, possible further research is discussed in the Conclusion.

Related Work

Many topic models have been proposed for topic evolution in the past decade. The key to these models lies in how to incorporate time into topic models. Depending on how they use time, most models can be generally classified into one of two categories.

The former category treats the document timestamp as an observed variable and uses a distribution over time to generate the timestamp (Dubey et al., 2013; Kawamae, 2011, 2012; Wang & McCallum, 2006). Among these models, TOT (Wang & McCallum, 2006) is a milestone and first models the words and timestamps jointly. TOT associates a beta distribution over time with each topic to capture its popularity. Kawamae (2011) proposed the trend analysis model (TAM) in which some topics are temporal and others are static. It more accurately captures temporal words and their changes. Dubey et al. (2013) proposed the nonparametric topic over time (npTOT) in which Gaussian mixture distribution is used to capture more flexible topic changes. Generally, these models are easy to extend, but it is difficult to choose a good distribution over time that allows both flexible changes and effective inferences.

DOI: 10.1002/asi

The latter category discretizes timestamps into epochs and models the evolution over these epochs (Ahmed & Xing, 2008, 2010; AlSumait, Barbará, & Domeniconi, 2008; Blei & Lafferty, 2006; Griffiths & Steyvers, 2004; Wang, Blei, & Heckerman, 2008). For example, Blei and Lafferty (2006) proposed the dynamic topic model (DTM), which uses a state space model on the natural parameters of the multinomial distributions that represent the topics. In contrast to the above models, the DTM is able to discover the evolving content of each topic. Based on the DTM, Wang et al. (2008) proposed continuous time DTM that uses Brownian motion to model topic changes. Thus, it does not require that time be discretized, and it is suitable for a sequential collection of documents. Ahmed and Xing (2008) proposed the recurrent Chinese restaurant process (RCRP) to model the birth, growth, and death of topics. Later work (Ahmed & Xing, 2010; Ahmed et al., 2011) based on the RCRP go further and help to organize documents effectively. These models capture topic changes in a more flexible manner and often simultaneously discover the changes of popularity and the drifts of content. However, time discretization plays an important role in the result. Improper time granularity either neglects fine changes or introduces noise.

Except for Apart from models in the above two categories, some other models can also be used for topic evolution. For instance, Blei and Frazier (2011) extended the traditional CRP to the distance-dependent Chinese restaurant process (ddCRP) to account for the inherent dependencies among data. It clusters dependent data into an infinite number of clusters. Kim and Oh (2011) proposed distance-dependent Chinese restaurant franchise (ddCRF) based on the table-based ddCRP, which is a mixture model. These distance-dependent models are capable of discovering topic evolution if the difference between timestamps is used to measure the distances between documents. An obvious advantage is that they are able to model various dependencies and flexible topic changes. Our work is based on the table-based ddCRP. We extend it to a nested process that can simultaneously model the dependencies among data and the relationship between clusters. Then HTEM is proposed to discover topic evolution in a hierarchy based on the nested process. Note that various timescales can be set in HTEM to decrease the influence of improper time granularity.

Furthermore, there are other models that are related to the proposed HTEM. First, to the best of our knowledge, two models have been proposed to discover the topic tree structure from dependent documents (Ahmed, Hong, & Smola, 2013; Nguyen, Boyd-Graber, & Resnik, 2013). Ahmed et al. (2013) proposed a hierarchical geographical model to model the user's locations and posts. It uses the nested Chinese restaurant franchise to model the tree structure and the multivariate Gaussian distribution to model the location. Nguyen et al. (2013) proposed supervised hierarchical latent Dirichlet allocation (SHLDA) to model the tree structure of topic words and response variables. The two models are easily extended for topic evolution, but limited in the form of the distribution that is used to generate the response variable. Comparing the two models, HTEM is able to capture more flexible changes. Second, several models that focus on multiscale topic evolution have been proposed (Iwata, Yamada, Sakurai, & Ueda, 2010; Nallapati, Ditmore, Lafferty, & Ung, 2007). For example, Nallapati et al. (2007) discretized time into epochs on multiple timescales, then discovered hierarchical topic changes. These models provide a way to zoom in and out on the timescale and study the evolution of topics at a chosen timescale. However, the topic structures in these models are generally constrained, and the consideration of the drift of content might not be suitable for clustering large document collections. Our HTEM captures the popularity changes of topics in a tree that can be infinitely branched and infinitely deep. The tree structure is very flexible and determined by real data.

Proposed Model

In this section, we propose a new topic model to discover hierarchical topic evolution in time-stamped documents. To introduce the proposed model clearly, we first review the ddCRP, and then show how to extend it to the nddCRP, which is able to model dependent data in a hierarchy. Next we propose the HTEM based on the new process to discover hierarchical topic evolution. We also develop a Gibbs sampler for the proposed HTEM.

Distance-Dependent Chinese Restaurant Process

The ddCRP is proposed to model data points with inherent dependencies (Blei & Frazier, 2011). It is able to model diverse kinds of data dependencies and cluster data points into an infinite number of clusters. However, there is no explicit cluster in the ddCRP, making it difficult to extend. Later, Kim and Oh (2011) modified the original ddCRP to the table-based ddCRP that directly assigns customers to tables.

In the table-based ddCRP, each customer entering the restaurant will choose a table on the basis of the cumulative decays between him and other customers who have already sat at the table. Let D denote the set of all distances between customers, d_{ij} denote the distance between customer i and j, z_i denote the index of the chosen table of ith customer, and K indicate the current number of tables. Let α be the scalar parameter and f be the decay function. Then the probability that ith customer will choose table z_i can be expressed as follows:

$$p(z_i = k \mid z_{1:(i-1)}, \alpha, f, D) \propto \begin{cases} \sum_{z_j = k} f(d_{ij}), & k \in K \\ \alpha, & k = K+1 \end{cases} \quad (1)$$

This process is denoted $ddCRP(\alpha, f, D)$. The scalar parameter α denotes the likelihood that a new customer might choose a new table. The decay function f mediates how the distances among customers affect the resulting distribution over partitions. There are three usually used decay functions: the windows decay $f(d) = 1[d < a]$, the exponential decay

DOI: 10.1002/asi

$f(d) = e^{-d/a}$, and the logistic decay $f(d) = \exp(-d + a)/(1 + \exp(-d + a))$. The distance set D contains a distance measure and all distances between each customer. Some popular distance measures include Euclidean distance, sequential distance, network hops, and so on. By setting different decay functions and distance measures, this process can be used to model various types of dependencies. For example, if the windows decay with a large parameter is used, this process is reduced to the traditional CRP. If the distance measure satisfies $d_{ij} = \infty$ for those $j > i$ and the decay function satisfies $f(\infty) = 0$, this process is equal to a sequential CRP.

In summary, the ddCRP utilizes the dependencies between data to provide a cluster bias that nearby data are more likely to be clustered in the same cluster. The ddCRP is flexible in applications with various distance measures and decay functions, and generally achieves a better performance than the traditional CRP.

Extending the ddCRP to Hierarchies

The ddCRP provides a flexible way to cluster dependent data into an infinite number of clusters. However, it ignores the relationship between clusters, which might lead to disordered clusters for large amounts of data. Several models have been proposed to model the relationship between clusters, among which Blei et al. (2010) introduced the nested Chinese restaurant process (nCRP). The nCRP organizes clusters in a hierarchy, providing a coarse-to-fine clustering.

The distance-dependent model and the nested model separately consider the dependencies among data points and the relationship between clusters. The two models are not conflicting and can even complement each other. Moreover, the dependencies among data and the relationship between clusters generally exist at the same time. Therefore, we extend the ddCRP to the nddCRP to organize dependent data in a hierarchy. In the nddCRP, there is a cluster tree that can be infinitely branched and infinitely deep. Each node in the tree corresponds to a cluster and also has a ddCRP distribution over its child nodes. Starting at the root node, each datum chooses a path to leaves based on these ddCRPs. Then the data are generated according to clusters in the chosen path. When assigning a table for a data point, it is important to consider the distances between the point and the points that have already been assigned to the table. Sometimes, a data point with close distance has more influence than several data points with far distance. Therefore, the nddCRP simultaneously models the relationship between clusters and the dependencies among data points.

Moreover, different measure scales can be set at different levels in the proposed nddCRP. Generally, large scales are set at a high level to capture the abstract dependencies, while small scales are set at the low level to model the specific dependencies. Each datum starts from the root, thus there is no need to set a measure scale in the first level. Then the measure scales vector has L-1 length for an L-depth tree, which is denoted $(s_1,\ldots,s_{L-1})$, and the resulting sets of distances can be represented as $(D_1,\ldots,D_{L-1})$. Therefore, the proposed nddCRP can be written as follows:

$$c \sim nddCRP(\alpha, f, D_{1:L-1}) \tag{2}$$

The nddCRP organizes dependent data in a hierarchy and models multiscale dependence. Moreover, it is a general model with various distance measures and decay functions that offers flexible and wide applications.

The Proposed Hierarchical Topic Evolution Model

The ddCRP has been used to discover topic evolution in time-stamped documents (Kim & Oh, 2011). Specifically, the dependence between time-stamped documents is measured by the difference between document timestamps. Documents with close timestamps are likely to talk about the same thing, whereas those with distant timestamps might focus on different things. After modeling topics, the popularity of each topic at each epoch is computed by the number of documents that generate words from this topic in this epoch. Topic evolution, such as birth, growth, and death, is also discovered from the popularity changes of topics.

Similarly, the nddCRP can also be used to discover topic evolution. It organizes evolving topics in a hierarchy. Each document is assigned to a path from the root to leaves in the tree and words in the document are generated from a distribution over topics in the chosen path. Note that the evolution of child topics constitutes the evolution of the parent topic in the nddCRP. Moreover, the nddCRP is able to discover multiscale topic evolution. Generally, the timestamps of documents are discretized into epochs at a given time granularity during the preprocessing (Ahmed et al., 2011; Blei & Lafferty, 2006). However, in our proposed nddCRP, different levels in the tree are able to be set with different time granularities. For example, at the second level the time granularity can be set as a year; at the third level the time granularity can be set as a month; and at the fourth level the time granularity can be even set as a day. Thus, topics at different levels of the tree are modeled to evolve at different timescales.

To formally describe the proposed hierarchical topic evolution model, we assume that each document (indexed by d) consists of a set of words w_d and a timestamp t_d. The nddCRP uses t_d to measure distances between documents and chooses a path c_d in the topic tree for each document. Then the words $w_{d,n}$ are generated from topics in the path c_d. This is a standard Dirichlet-Multinomial framework. Let $z_{d,n}$ denote the index of level for word $w_{d,n}$. The generative process of the proposed HTEM is represented in Figure 1 and defined as follows:

1. For each table $k \in$ T in the infinite tree,
   a. Draw a topic $\phi_k \sim Dirichlet(\eta)$.
2. For each document, $d \in \{1,2,\ldots,M\}$
   a. Draw a path $c_d \sim nddCRP(\alpha, f, D_{1:L-1})$.
   b. Draw a distribution over levels in the tree $\theta_d \mid \{m, \pi\} \sim GEM(m, \pi)$.
   c. For each word,
      i. Choose level $z_{d,n} \mid \theta_d \sim Discrete(\theta_d)$.

DOI: 10.1002/asi

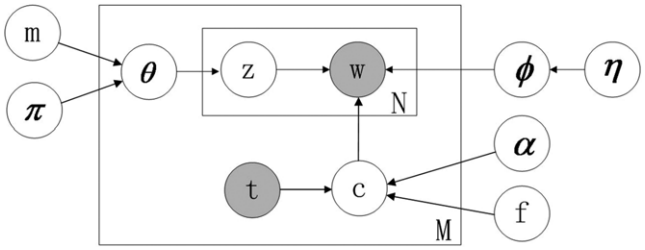

FIG. 1. Graphical model of hierarchical topic evolution model.

ii. Choose word $w_{d,n} \mid \{z_{d,n}, c_d, \phi\} \sim Discrete(\phi_{c_d[z_{d,n}]})$, which is parameterized by the topic in level $z_{d,n}$ in the path c_d.

We use GEM distribution as the prior distribution on the levels. The GEM distribution has two hyperparameters, the scaling parameter π that determines the confidence of the prior and the ratio parameter m that represents the ratio to stay at the current node. By using a stick-breaking construction, the GEM provides a distribution over infinite partitions of a unit interval. Specifically, view the interval (0, 1) as a unit-length stick. Draw a value V_1 from a $Beta(m\pi, (1-m)\pi)$ distribution and break off a fraction V_1 of the stick. Then the first fragment of the stick is denoted $\theta_1 = V_1$. Continue this procedure recursively for the remainder of the stick and the second fragment of the stick is denoted $\theta_2 = V_2(1-\theta_1)$. In general, the fragments of the stick are defined as follows:

$$\theta_i = V_i \prod_{j=1}^{i-1}(1-V_j), \quad where \quad V_i \sim Beta(m\pi, (1-m)\pi) \quad (3)$$

Therefore, the GEM distribution makes the hierarchy infinitely deep and generally higher probabilities are allocated to topics at the higher levels. Note that if we truncate the stick-breaking process after $L-1$ breaks, we obtain a Dirichlet distribution on an L-dimensional vector.

Compared with the hierarchical latent Dirichlet allocation (HLDA) (Blei et al., 2010), the proposed HTEM simultaneously models the relationship between topics and the dependencies among documents. It not only organizes evolving topics in a tree but also makes higher-level topics be more abstract and evolve more slowly.

Posterior Inference

Several methods have been developed to estimate latent variables in a probabilistic graphical model and Gibbs sampling generally yields a relatively simple algorithm. The main idea of Gibbs sampling is to sample a new value for a variable according to its conditional probability given other variables. We describe a Gibbs sampling method for the proposed HTEM. For simplicity, we fix the hyperparameters, including m, π, η, and α. Moreover, the word distribution ϕ and the topic distribution θ are obtained directly by the conjugate Dirichlet-Multinomial distribution. Therefore, what we need to sample is the path of each document and the level assignment of each word in each document.

Sampling path. The path of a document is represented as a series of topic nodes from the root to the leaves in the hierarchy. When sampling the path, each document either chooses an existing path in the hierarchy or creates a new path from an internal topic node. Because of the exchangeability of child nodes, possible paths are finite for a document. Therefore, computing the conditional probabilities of all possible paths for the document is feasible. Given all level assignments and paths of other documents, the detailed conditional probability of a path for a document is defined as:

$$\begin{aligned} &p(c_d \mid w, c_{-d}, z, \eta, \alpha, f, D_{1:L-1}) \\ &\quad \propto p(c_d \mid c_{-d}, \alpha, f, D_{1:L-1}) p(w_d \mid c, w_{-d}, z, \eta) \end{aligned} \quad (4)$$

where the former part is the prior of the nddCRP that is computed as Equation 1 and the sets of distances $D_{1:L-1}$ are precomputed with document timestamps before sampling. The latter part is the likelihood of the words. Given a chosen path, the likelihood is a standard Dirichlet-Multinomial integral and can be simplified as:

$$\begin{aligned} &p(w_d \mid c, w_{-d}, z, \eta) \\ &\quad = \prod_{l=1,k=c_{d,l}}^{L} \frac{\prod_v \Gamma(n_{k,v} + \eta)}{\prod_v \Gamma(n_{k,v\backslash d} + \eta)} \frac{\Gamma\left(\sum_v n_{k,v\backslash d} + V\eta\right)}{\Gamma\left(\sum_v n_{k,v} + V\eta\right)} \end{aligned} \quad (5)$$

in which $n_{k,v}$ is the number of times that word v is assigned to topic k, and $n_{k,v\backslash d}$ is the same number except in document d. V is the vocabulary size and η is the parameter of Dirichlet prior. L is the max depth of the topic tree.

Sampling level. After sampling a path for a document, each word samples a level assignment to choose a topic in the path to generate the word. For each word, the conditional probability of the level assignment given the path is defined as:

$$\begin{aligned} &p(z_{d,n} \mid z_{-(d,n)}, c, w, m, \pi, \eta) \\ &\quad \propto p(z_{d,n} \mid z_{d,-n}, m, \pi) p(w_{d,n} \mid z, c, w_{-(d,n)}, \eta) \end{aligned} \quad (6)$$

where the first part is the posterior of the GEM distribution that can be computed as:

$$p(z_{d,n} = k \mid z_{d,-n}, m, \pi) = \frac{m\pi + n_{d,k\backslash n}}{\pi + N_{d,k\backslash n}} \prod_{j=1}^{k-1} \frac{(1-m)\pi + N_{d,j+1\backslash n}}{\pi + N_{d,j\backslash n}} \quad (7)$$

where $n_{d,k\backslash n}$ is the number of times that words in document d are assigned with level k except the nth word. The $N_{d,k\backslash n}$ is defined as $\sum_{i=k}^{L} n_{d,i\backslash n}$, which is the number of times that words in document d are assigned to levels larger or equal to k except the nth word.

The second part is the likelihood of a word for the topic. It is simply the probability of the current word in the given topic, which is computed as:

DOI: 10.1002/asi

$$p(w_{d,n} \mid z, c, w_{-(d,n)}, \eta) = \frac{n_{k,w_{d,n}} + \eta}{\sum_v (n_{k,v} + \eta)} \quad (8)$$

where $n_{k,v}$ is the number of times that word v is assigned to topic k.

Given paths, the level assignments are sampled based on Equation 6 for each word in the document, and then given the level assignments, the paths are sampled based on Equation 4 for each document. The two sampling processes are iterated until convergence. Eventually, documents are assigned to paths of the topic tree and document words are generated from topics. Then the topics and their evolutions are inferred from these latent variables.

Experiments

In this section, we apply the proposed HTEM to discover hierarchical topic evolution on two real-world data sets. We first use the HTEM to discover the topic tree from the two data sets and then visualize the hierarchical evolution in a timeline. We further show that the proposed HTEM discovers more coherent topics and obtains higher held-out likelihood than the baseline models.

Data Description

The time span of a data set is of great importance to test topic evolution models. Therefore, we examine the proposed HTEM on two real-world data sets, among which one has a large time span over 200 years and the other has a very small time span of just more than 10 years. Moreover, their contents are in different domains. The two data sets are described in detail as follows.

State of the Union Addresses. The State of the Union Addresses data set reflects the history of the United States and consists of 214 transcripts presented by the president to congress from 1790 to 2006. Following Wang and McCallum (2006) in preprocessing the data set, we split each transcript into three-paragraph documents, ignore case, and remove stop words and words appearing less than 10 times in the data set. Documents from the same transcript have the same timestamp when the address was given.

Neural Information Processing Systems [NIPS] Papers. The NIPS Papers data set contains the full text of the 13 years of proceedings from 1987 to 1999 of NIPS conferences. The preprocessed data set is available online (Roweis, n.d.[1]). Each document's timestamp is the year when the paper was published. The detailed statistics of the two data sets are presented in Table 1.

TABLE 1. Summary statistics of two experiment data sets.

Data set	#documents	#unique words	#word tokens	Timestamp
Addresses	6,499	7,424	757,080	1790–2006
NIPS	1,740	13,649	3,267,278	1987–1999

Topic Tree Discovery

In this part, we employ our HTEM to discover the topic tree from the two data sets. If the discovered tree organizes topics reasonably and the extracted topics are in accord with real topics, both in semantics and active time, it will be proven that the HTEM is effective to discover hierarchical topic evolution.

For both data sets, the stick-breaking procedure in the GEM is truncated at three levels for simplicity and the logistic decay function is chosen. The time granularities are set according to the time span of the data set. A good setting of time granularities results in proper number of epoch at each level of the tree. Based on the time span for the Addresses data set, time granularities are set as 6 years at the second level and 2 at the third level, whereas for the NIPS data set, time granularities are set as 2 years at the second level and 1 at the third level. Documents in the NIPS data set are longer in length and so higher topic prior parameters and the GEM parameters are set for the data set. The detailed parameter settings are shown in Table 2. We rerun each model 10 times and take the result with the mean score as the final result. For each run, the burn-in is set as 200, the sample lag is set as 1, and the total iteration is set as 2,000.

The discovered topic trees are shown in Figures 2 and 3. Each topic tree is too large to fit in the space provided, so we take a subtree to illustrate the discovered topic tree. Each topic node in the tree contains a plot of popularity changes over time and the top 10 words with highest probability. The topic popularity is measured by the number of documents that choose the topic to generate their words. Moreover, each topic is named according to the semantics of the top words for easy reading.

Figure 2 shows the topic tree discovered from the Addresses data set. At the first level, it is the background topic that includes the most common words used in the addresses, such as "government," "states," "congress," and "great." At the second level, there are four abstract topics, "Treaty," "Civil Rights," "Military," and "Social Development" topic. The four topics reflect major issues in American history that existed for a long period. At the third level, each abstract topic has several child topics that generally evolve in a short timescale and coincide with real events in history. For example, under the "Civil Rights" topic there are two child topics, the announcement of the Emancipation Proclamation in 1862 that abolished slavery in the United States and the promulgation of the Dawes Act in 1887 that distributed land to Indians, both of which are remarkable events in the development of American civil rights. As for the "Military" topic, it is further described by three wars: the Spanish-American War, the Second World War, and the Iraq War. The three wars are very significant

[1]http://www.cs.nyu.edu/~roweis/data.html

DOI: 10.1002/asi

TABLE 2. Parameter settings in experiments.

Data set	Time granularities	Topic prior parameters	GEM parameters	Decay parameters	Scaling parameter
Addresses	6, 2	0.5, 0.3, 0.1	20, 0.5	3	1e-6
NIPS	2, 1	1.0, 0,7, 0.5	200, 0.5	2	1e-8

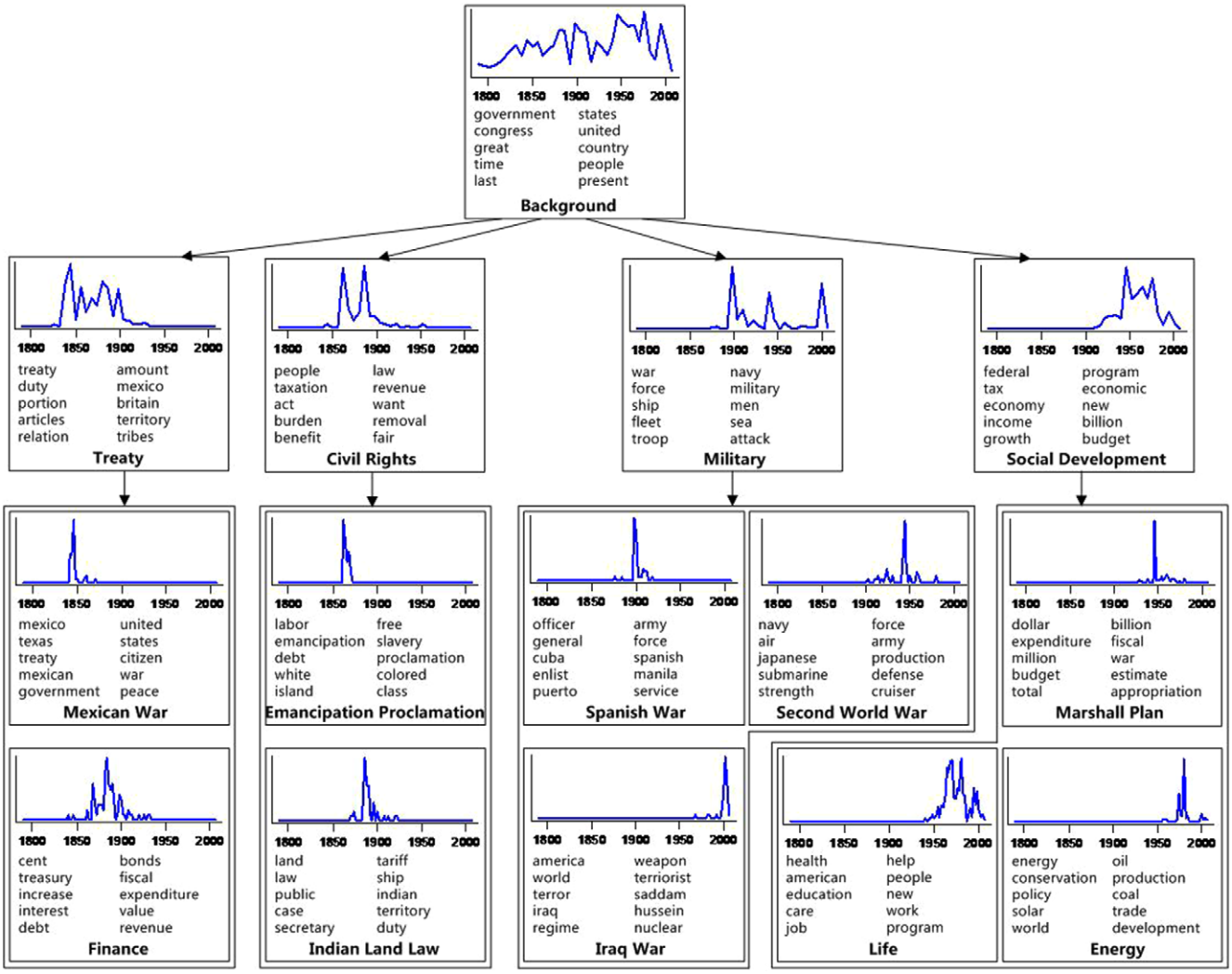

FIG. 2. Topic tree discovered from the Addresses data set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in American military history and describe the military development in different periods. From the discovered topic tree, American history is presented from abstract topics to specific topics, which is helps the user to understand the entire American history.

Figure 3 presents the topic tree discovered from the NIPS data set. The background topic at the first level indicates that the NIPS conference mainly focuses on neural network and model learning by top words such as "network," "learning," "neural," and "model." The abstract topics at the second level reveal some hot research directions at NIPS conferences, including "Optimization Algorithm," "Transformation," "Parameter Estimation," and "Classification" topics. At the third level, the "Transformation" topic has two child topics, the "Tangent Distance" topic and the "ICA" topic. The tangent distance is a new transformation invariant distance measure proposed in 1993 and the independent components analysis (ICA) is a special linear transformation proposed in 1994. Besides, the "Classification" topic consists of two child topics, the "SVM" topic and the "Boost" topic. Both of them are still very important classification methods. The discovered topic tree clearly shows the important scientific research hotspots and their evolutions.

The HTEM is not only capable of accurately extracting topics from time-stamped documents but also able to

DOI: 10.1002/asi

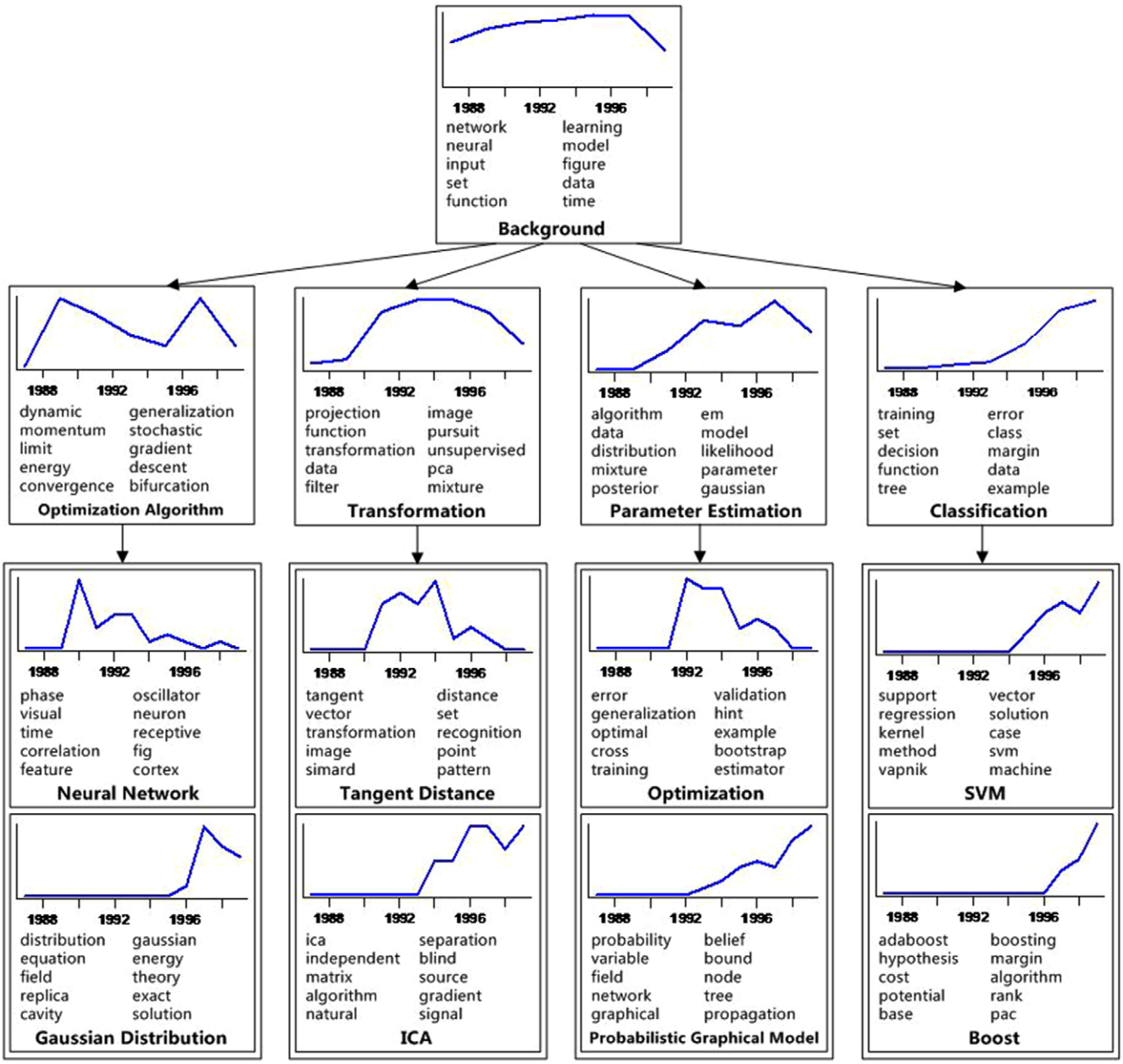

FIG. 3. Topic tree discovered from the NIPS data set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

explicitly discover their hierarchical evolution in a tree. Topics at the low level are more specific and evolve in a shorter timescale than topics at the high level. The resulting tree structure provides coarse-to-fine information retrieval, which is useful and effective for document analysis and organization.

Figure 4 illustrates the log likelihood as a function of sampling iteration. Iterations 10–1,000 are plotted for a clearer figure. It shows that the HTEM finally converges in the iteration on both data sets. Note that the HTEM converges faster on the NIPS data set than on the Addresses data set. Among the two data sets, documents in the Addresses data set are shorter in length and their semantics are more ambiguous. Thus, it needs more iteration to converge. Generally, the HTEM Gibbs sampler is effective for estimating the model.

Hierarchical Evolution Visualization

Topics in the tree can be born, grow, and die as time goes on. Visualizing this process can describe the topic evolution dynamically and intuitively. In this part, we visualize the discovered topic tree by the HTEM in timeline.

The birth and death time of each topic are important for topic evolution visualization. The popularity changes of

DOI: 10.1002/asi

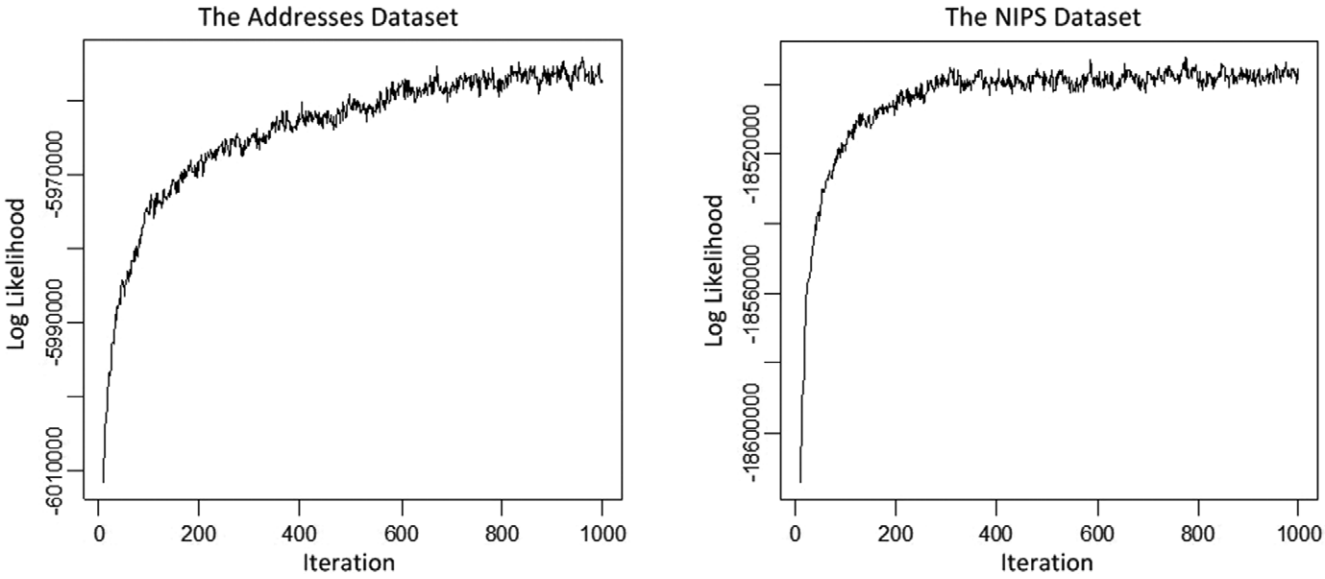

FIG. 4. The log likelihood over the iteration.

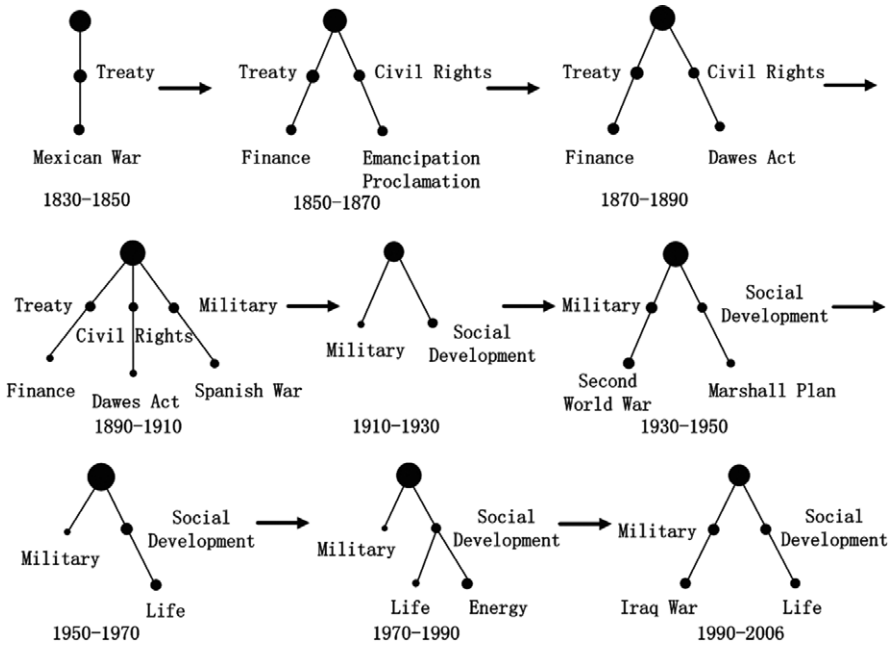

FIG. 5. A visualization of hierarchical topic evolution for the Addresses data set (the size of the circle indicates the topic's intensity).

topics discovered by HTEM might contain noise, so it is less accurate to take the first and last timepoint with nonzero intensity as the birth and death time. Some simple methods can help to determine the time. For example, the birth time can be simply set as the timepoint when the cumulative intensity reaches a certain proportion of the total intensity or when the intensity first achieves a given percentage of the maximum intensity. Here we combine these ideas to reduce the influence of noise. Then according to the birth and death time of topics, we can plot the tree structure for each epoch and visualize its evolution in a timeline. Figure 5 shows the hierarchical evolution for the Addresses data set. For space consideration, we show the hierarchical changes for each 20 years, and each hierarchical tree is a subtree of the global hierarchical tree. Only topic nodes in Figure 2 are shown and the size of the circle in the figure is proportional to topic intensity.

Figure 5 presents the evolution of the hierarchical tree structure over time. Topics in the tree emerge, grow, and disappear as time goes on. For example, the "Mexican War"

DOI: 10.1002/asi

topic emerges in the first epoch (1830–1850) and disappears quickly, whereas "Civil Rights" is a general topic and continues to evolve between 1850 and 1910. From its child topics, we know that the development of American civil rights is first promoted by the Emancipation Proclamation, and then it reaches another climax with the Dawes Act. As for the "Military" topic, it exists and evolves over a long period. Some child topics emerge and disappear successively over the course of its development, such as the "Spanish-American War" topic in the 4th epoch, the "Second World War" topic in the 6th epoch, and the "Iraq War" topic in the last epoch. The change of these child topics quickly and concisely reflects the evolution of the American military. In general, hierarchical evolution visualization dynamically shows the changes in American history, and is very clear and intuitive.

Therefore, from the hierarchical topic evolution discovered by the HTEM, topics can be shown in a tree and dynamically in a timeline from abstract to specific. The HTEM provides a new, convenient way for information retrieval.

Topic Analysis

A topic is represented by a distribution over words. Here, we analyze the semantics of discovered topics from the HTEM and the baseline model, that is, ddCRF (Kim & Oh, 2011). Based on the table-based ddCRP, the ddCRF is a flat mixture model while the HTEM is a hierarchical model. By comparing the discovered topics from ddCRF and the HTEM, we can measure the improvement from adding hierarchy.

As with the HTEM, we choose the logistic decay function with the same parameter in ddCRF. The table scaling parameter and dish scaling parameter are set as 0.1. The topic prior parameter is set as 0.1 for the Addresses data set and 0.2 for the NIPS data set to obtain a similar count of topics. Some topics are simultaneously discovered from the HTEM and ddCRF and we show them in Figure 6.

In Figure 6, each topic is shown by its popularity changes over time and top words with their probabilities. It is easy to see that the HTEM discovers more meaningful topics. For example, in the first topic about the Emancipation Proclamation, although the popularity change over time obtained by the two models are similar, the HTEM captures more specific words than ddCRF, which contains some abstract words such as "people," "union," and "country." The second is an abstract topic about the military. Both models capture abstract topic words and multimodal popularity change. Note that the topic of the HTEM has some child topics to specific meanings, which is shown in Figure 2. However, the topic of ddCRF has not. About the last two topics, they are discovered from the NIPS data set. Figure 3 shows some abstract words about the NIPS data set in the root topic, such as "network," "learning," "neural," "data," and "function." We find that topics discovered by the HTEM about tangent distance and SVM hardly contain these abstract words. Thus, they have clearer meanings than those of ddCRF. Moreover, the topic popularity change of the HTEM more precisely matches the true situation. Therefore, topics discovered by the HTEM are more specific and meaningful than those by ddCRF.

Generally, the hierarchical model makes topics at the high level more abstract and those at the low level more specific, and presents topics from coarse to fine. Moreover, topics discovered by HTEM are more coherent and meaningful than the flat model.

Held-Out Likelihood

Held-out likelihood is widely used in topic models to evaluate how well the trained model explains the held-out data (Griffiths & Steyvers, 2004; Wallach, Murray, Salakhutdinov, & Mimno, 2009). Generally, a better model will give a higher held-out likelihood. Given a trained model, the held-out likelihood is calculated as:

$$Heldout\ likelihood = \log p(W \mid M_{train}) \tag{9}$$

where M_{train} denotes the trained model and W denotes the held-out data. In this experiment, we use the first 90% of documents for training and the rest for testing. For an overall comparison, we evaluate our model with various parameter settings of the exponential and logistic decay functions. We also compare our model with the baseline models, including ddCRF and HLDA.

Figure 7 illustrates the held-out likelihood of HTEM, ddCRF, and HLDA on the two data sets. The hyperparameters of the HTEM are shown in Table 2 except the decay parameter. The hyperparameters of ddCRF are set according to the HTEM to obtain an equivalent number of topics to the HTEM. The sampling rate of the test data is set as 10. For both data sets, the HTEMs with logistic and exponential decay function generally achieve similar held-out likelihoods, which are higher than ddCRFs. Moreover, the HTEMs on average achieve better held-out likelihood than HLDA, which is a special model of the HTEM. When the decay parameter becomes larger, the held-out likelihood of the HTEM is generally closer to the value of HLDA. Therefore, the figure proves that extending the model to hierarchy is necessary and effective.

The held-out likelihood experiment verifies that models trained by the HTEM explain held-out data better than ddCRF and HLDA. Moreover, the HTEM with lower decay parameters is more sensitive to the distance and various values can be set for different applications.

Conclusion

In this paper, we propose a new topic model to discover hierarchical topic evolution in time-stamped documents. We firstly extend the ddCRP to a nested process to model dependent data in a hierarchy, and then propose our HTEM based on the extended process for hierarchical topic evolution. In

DOI: 10.1002/asi

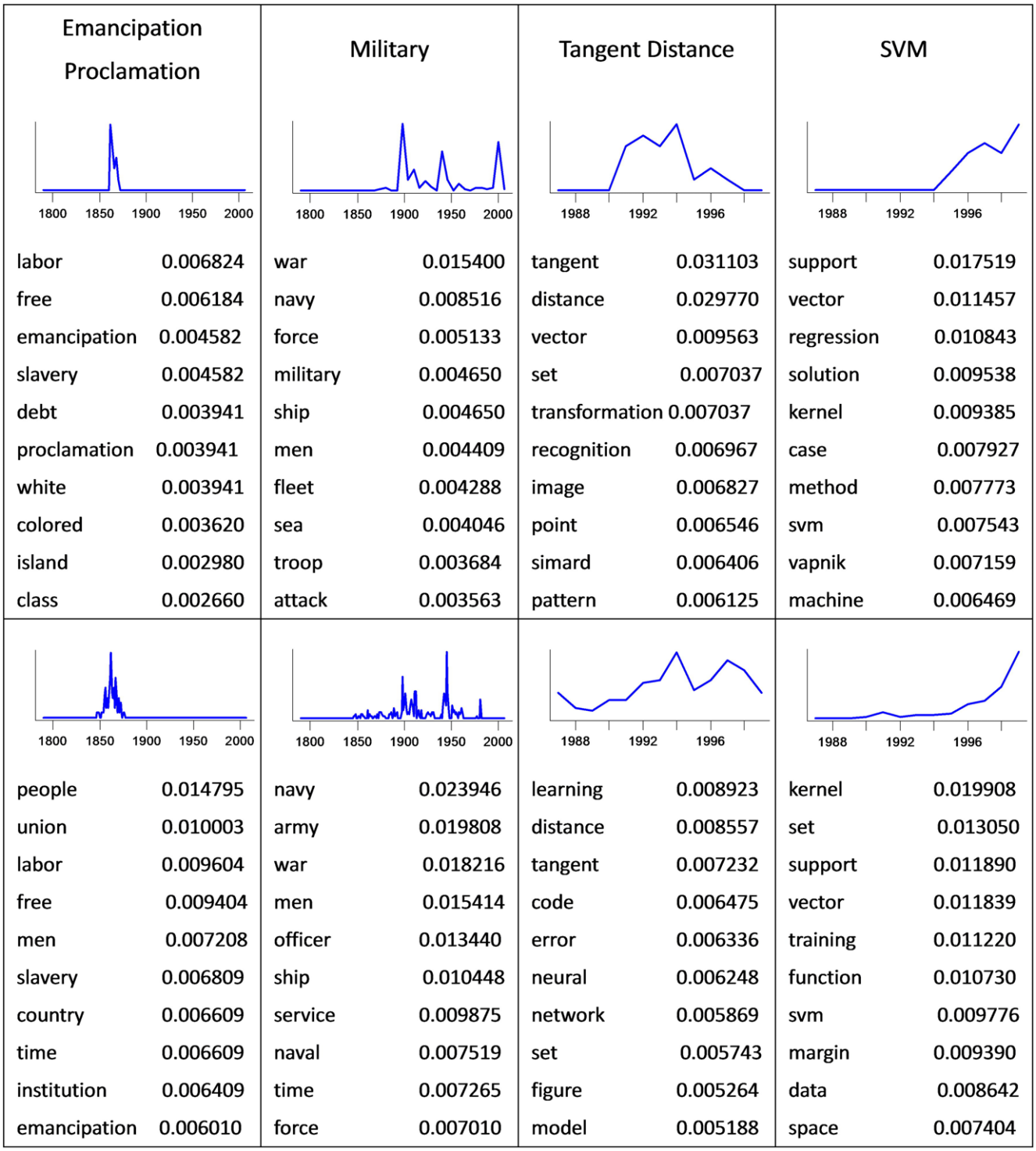

FIG. 6. Four topics discovered by HTEM (above) and ddCRF (bottom). The left two columns are topics discovered from the Addresses data set and the right two are mined from the NIPS data set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the experiment, we apply the proposed HTEM on two real-world data sets, the State of the Union Addresses data set and the NIPS papers data set.

The experimental results show that the topic tree discovered from the Addresses data set reflects some abstract topics about American history at the high level, such as civil rights, military, and social development. At the low level, for example, the military topic has three specific child topics: the Spanish-American War, the Second World War, and the Iraq War. The topic tree discovered from the

DOI: 10.1002/asi

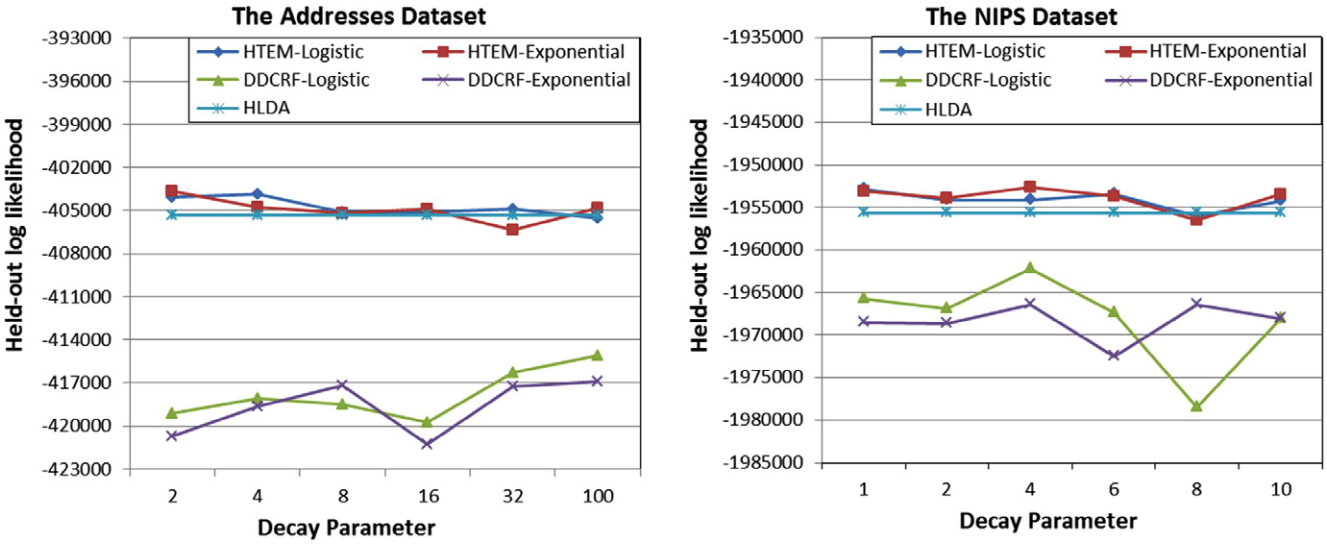

FIG. 7. Held-out likelihood of the two experiment data sets. (Higher is better. The HTEM outperforms ddCRF and HLDA.) [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

NIPS data set shows some scientific research hotspots, including optimization algorithm, transformation, parameter estimation, and classification. And the classification topic also has two child topics, SVM and boost. And last, the HTEM achieves a better held-out likelihood than the baseline on both data sets. Therefore, all the experimental results verify that the proposed HTEM is effective at discovering hierarchical topic evolution in time-stamped documents.

The proposed HTEM can be extended in at least two ways. On one hand, it can be used to organize and analyze other dependent data with different distance measures, such as geo-tagged data (Ahmed et al., 2013), linked data (Ho, Eisenstein, & Xing, 2012), and image data (Ghosh, Ungureanu, Sudderth, & Blei, 2011). On the other hand, it can be extended with other models, including other hierarchical models, online models, and evolution models for analyzing the drifts of content. The results of these ongoing research works will be reported in the near future.

Acknowledgments

This research is supported by National High Technology Research and Development Program of China (Grant No. 2012AA011005).

References

Ahmed, A., & Xing, E.P. (2008). Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: With applications to evolutionary clustering. Paper presented at the SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). Atlanta, Georgia, April 2008.

Ahmed, A., & Xing, E.P. (2010). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. Paper presented at the Proceedings of the 26th International Conference on Conference on Uncertainty in Artificial Intelligence. Corvallis, OR: AUAI Press. Catalina Island, California, July 210.

Ahmed, A., Ho, Q., Teo, C.H., Eisenstein, J., Smola, A.J., & Xing, E.P. (2011). Online inference for the infinite topic-cluster model: Storylines from streaming text. Paper presented at the International Conference on Artificial Intelligence and Statistics. Ft. Lauderdale, FL, USA, April 2011.

Ahmed, A., Hong, L., & Smola, A.J. (2013). Hierarchical geographical modeling of user locations from social media posts. Paper presented at the Proceedings of the 22nd International Conference on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Rio de Janeiro, Brazil, May 2013.

AlSumait, L., Barbará, D., & Domeniconi, C. (2008). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. Paper Presented at the Proceeding of the 8th IEEE International Conference on Data Mining. Los Alamitos, CA: IEEE Computer Society. Pisa, Italy, December 2008.

Blei, D.M., & Frazier, P.I. (2011). Distance dependent Chinese restaurant processes. Journal of Machine Learning Research, 12, 2461–2488.

Blei, D.M., & Lafferty, J.D. (2006). Dynamic topic models. Paper presented at the Proceedings of the 23rd International Conference on Machine Learning. New York: ACM. Pittsburgh, Pennsylvania, June 2006.

Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 1(1), 17–35.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Blei, D.M., Griffiths, T.L., & Jordan, M.I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM, 57(2), 1–30.

Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., & Uyar, E. (2010). New event detection and topic tracking in Turkish. Journal of the American Society for Information Science and Technology, 61(4), 802–819.

Dubey, A., Hefny, A., Williamson, S., & Xing, E.P. (2013). A non-parametric mixture model for topic modeling over time. Paper presented at the SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). Texas, USA, May 2013.

Ghosh, S., Ungureanu, A.B., Sudderth, E.B., & Blei, D.M. (2011). Spatial distance dependent Chinese restaurant processes for image segmentation. Paper presented at the Advances in Neural Information Processing Systems. La Jolla, CA: Neural Information Processing Systems Foundation. Neural Information Processing Systems Foundation.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl. 1), 5228–5235.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? Paper presented at the Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM. Hong Kong, China, November 2009.

Ho, Q., Eisenstein, J., & Xing, E.P. (2012). Document hierarchies from text and links. Paper presented at the Proceedings of the 21st International Conference on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Lyon, France, April 2012.

Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2010). Online multiscale dynamic topic models. Paper presented at the Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM. Washington, DC, July 2010.

Kawamae, N. (2011). Trend analysis model: Trend consists of temporal words, topics, and timestamps. Paper presented at the Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. New York: ACM. Hong Kong, China, February 2011.

Kawamae, N. (2012). Theme chronicle model: Chronicle consists of timestamp and topical words over each theme. Paper presented at the Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM. Sheraton, Maui Hawaii, November 2012.

Kim, D., & Oh, A. (2011). Accounting for data dependencies within a hierarchical dirichlet process mixture model. Paper presented at the Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM. Glasgow, Scotland, UK, October 2011.

Lin, N., Li, D., Ding, Y., He, B., Qin, Z., Tang, J., & Dong, T. (2012). The dynamic features of Delicious, Flickr, and YouTube. Journal of the American Society for Information Science and Technology, 63(1), 139–162.

Nallapati, R.M., Ditmore, S., Lafferty, J.D., & Ung, K. (2007). Multiscale topic tomography. Paper presented at the Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM. San Jose, CA, August 2007.

Nguyen, V.-A., Boyd-Graber, J., & Resnik, P. (2013). Lexical and hierarchical topic regression. Paper presented at the Advances in Neural Information Processing Systems. La Jolla, CA: Neural Information Processing Systems Foundation. Lake Tahoe, Nevada, United States, December 2013.

Sugimoto, C.R., Li, D., Russell, T.G., Finlay, S.C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American Library and Information Science dissertations using latent Dirichlet allocation. Journal of the American Society for Information Science and Technology, 62(1), 185–204.

Wallach, H.M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM. Montreal, Quebec, June 2009.

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous time dynamic topic models. Paper presented at the 23rd Conference on Uncertainty in Artificial Intelligence. Corvallis, OR: AUAI Press. Helsinki, Finland, July 2008.

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. Paper presented at the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM. Philadelphia, USA, August 2006.

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. Journal of the American Society for Information Science and Technology, 63(8), 1631–1646.

DOI: 10.1002/asi