Extraction of Topic Evolutions from References in Scientific Articles and Its GPU Acceleration

Tomonari Masada  
Nagasaki University  
1-14 Bunkyo-machi, Nagasaki-shi  
Nagasaki, 852-8521 Japan  
masada@nagasaki-u.ac.jp

Atsuhiro Takasu  
National Institute of Informatics  
2-1-2 Hitotsubashi, Chiyoda-ku  
Tokyo, 101-8430 Japan  
takasu@nii.ac.jp

ABSTRACT
This paper provides a topic model for extracting topic evolutions as a corpus-wide transition matrix among latent topics.

Recent trends in text mining point to high demand for exploiting metadata, i.e., data about data. Especially, exploitation of reference relationships among documents induced by e.g. hyperlinking Web pages, citing scientific articles, tumbling blog posts, retweeting tweets, etc, is put in the foreground of the effort to achieve an effective mining. We focus on scholarly activities and propose a new topic model for obtaining a corpus-wide view on how research topics evolve along reference relationships among scientific articles. Our topic model, called TERESA, extends latent Dirichlet allocation (LDA) by introducing a corpus-wide topic transition probability matrix, with which we model reference relationships among articles as transitions between latent topics. Our approximated variational inference updates LDA posteriors and topic transition posteriors alternately. The main issue is execution time, because the time complexity is proportional to $MK^2$, where $K$ is the number of topics and $M$ is the number of reference relationships (i.e., links in citation network). Therefore, we accelerate the inference with Nvidia CUDA. We prove the effectiveness of TERESA by introducing a new evaluation measure called diversity plus focusedness $(D+F)$. We also present examples of topic evolutions extracted by our method.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—Data Mining; I.2.6 [Artificial Intelligence]: Learning

General Terms
Algorithms, Experimentation

Keywords
Topic modeling, Citation analysis, GPU

1. INTRODUCTION
Intensive analysis of reference relationships among documents is a key factor to success in today’s text mining. In fact, this type of relationship prevails. For example, hyperlinking Web pages, citing academic articles, tumbling blog posts, retweeting tweets, etc, induce such relationships among documents. Classic link analysis [17, 9] puts less emphasis on text data found in each document. In contrast, we adopt LDA [4], a promising framework for effective text mining, and extend it for exploiting reference relationships.

In this paper, we focus on academic activities and propose a new LDA-like topic model for analyzing reference relationships among scientific articles. Since our model extracts Topic Evolutions from REferences in Scientific Articles, we call it TERESA by concatenating the italicized capital letters. The most important outcome obtained by TERESA is a corpus-wide topic transition matrix whose $(i,j)$th entry gives the posterior transition probability from topic $t_i$ to topic $t_j$. Roughly, when the articles containing topic $t_i$ with a high probability are often cited by the articles containing topic $t_j$ with a high probability, the $(i,j)$th entry of this matrix gives a high transition probability. We can interpret this matrix as a summary of frequent topic evolutions latent in an article set. We call a task to extract a set of corpus-wide topic transition patterns topic evolution analysis.

The main contributions of this paper are as follows:

1. This paper proposes a method for topic evolution analysis, where we can obtain a corpus-wide and directed relationships among topics. First, TERESA provides a corpus-wide view on how research topics evolve along reference relationships. Our aim is to obtain an accumulation of local views, e.g. a series of topic transition patterns between two adjacent time points along the time axis [3, 18]. This type of accumulation of local views makes it difficult to get a global view on how topics interrelate. Second, TERESA extracts directed relationships among topics. Therefore, we can respect the direction intrinsic to every citation, though some preceding works neglect the direction by regarding each relationship as symmetric [15, 5]. For evaluation, we devise a new measure called diversity plus focusedness $(D+F)$ and compare TERESA with LDA.

2. This paper proposes an acceleration of inference for TERESA. We employ variational Bayesian inference (VB) for obtaining an estimation of the posterior parameters. In our VB, a large number of posterior parameters can be updated independently. Therefore,
2. RELATED WORKS

In social media, we meet a wide variety of interactions among entities, e.g., persons, documents, products, places, images, movies, etc. Compelling data mining methods are expected to exploit such relationships observable among various types of entities. This paper focuses on reference relationships among scientific articles, because this type of relationship may help us to find papers we can cite and to predict what kind of research will follow ours.

Many methods are proposed to use reference relationships among scientific articles for topic extraction. However, our method is different from them as is discussed below.

2.1 Model Complexity

Nallapati et al. [14] propose a method for modeling topic transitions among documents. While we can find works that do not give any direct modelings of relationships of topic distributions between cited and citing documents [13, 7, 11], Nallapati et al. explicitly model transitions of topic distributions among documents as “flows” of topics. However, their probabilistic model has $\mathcal{K}$ parameters at every link in citation network. Therefore, the complexity of the parameters for modeling citation links amounts to $O(M\mathcal{K})$, where $\mathcal{K}$ is the number of topics and $M$ is the number of references, i.e., the number of links in citation network. It can be said that this model aims at giving an accumulation of local transition patterns, and thus that it is difficult to obtain a global view on how topics are related to each other. We may repeat the same discussion with respect to [19], where relationships of topic distributions among documents are modeled as Markov random fields. Further, Ren et al. [18] modifies HDP [20] so that we can distill topic transitions between adjacent time points along the time axis. That is, they consider topic transitions not between linked documents as in [14] but between adjacent time points. Therefore, the complexity of the parameter space is reduced. However, this proposal also makes us uncomfortable with respect to the issue how we can integrate a pile of time-dependent topic transitions to obtain a unified view on how topics are related to each other. We may repeat the same discussion with respect to [19], where relationships of topic distributions among documents are modeled as Markov random fields.

In contrast, TERESA extracts a single $\mathcal{K} \times \mathcal{K}$ matrix of topic transition probabilities and thus has a smaller number of parameters for modeling citation network of scientific articles. In this manner, we obtain a global view on how topics are related to each other. We may repeat the same discussion with respect to [19], where relationships of topic distributions among documents are modeled as Markov random fields. Further, Ren et al. [18] modifies HDP [20] so that we can distill topic transitions between adjacent time points along the time axis. That is, they consider topic transitions not between linked documents as in [14] but between adjacent time points. Therefore, the complexity of the parameter space is reduced. However, this proposal also makes us uncomfortable with respect to the issue how we can integrate a pile of time-dependent topic transitions to obtain a unified view on how topics are related to each other. A similar discussion may also be valid for [3, 22].

In contrast, TERESA extracts a single $\mathcal{K} \times \mathcal{K}$ matrix of topic transition probabilities and thus has a smaller number of parameters for modeling citation network of scientific articles. In this manner, we obtain a single corpus-wide view of topic interaction as a digraph whose vertices are latent topics and whose weighted arcs are transitions among topics accompanied with their probabilities (Figure 1). Therefore, the result of our topic evolution analysis is easy to grasp.

2.2 Computation Acceleration
Table 1: Definitions of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D = {d_1, \ldots, d_J} )</td>
<td>set of documents</td>
</tr>
<tr>
<td>( \mathcal{V} = {v_1, \ldots, v_W} )</td>
<td>set of words</td>
</tr>
<tr>
<td>( T = {t_1, \ldots, t_K} )</td>
<td>set of latent topics</td>
</tr>
<tr>
<td>( \mathcal{C}<em>j = {d</em>{j1}, \ldots, d_{jC_{C_j}}} )</td>
<td>set of documents cited by document ( d_j )</td>
</tr>
<tr>
<td>( x_j = {x_{j1}, \ldots, x_{jK}} )</td>
<td>word tokens in document ( d_j )</td>
</tr>
<tr>
<td>( z_j = {z_{j1}, \ldots, z_{jK}} )</td>
<td>topic assignments in document ( d_j )</td>
</tr>
<tr>
<td>( \theta_j = (\theta_{j1}, \ldots, \theta_{jK}) )</td>
<td>( \theta_{jk} ), probability that a word token in document ( d_j ) is assigned to topic ( t_k )</td>
</tr>
<tr>
<td>( \phi_k = (\phi_{k1}, \ldots, \phi_{kW}) )</td>
<td>( \phi_{kw} ), probability that a token of word ( v_w ) is assigned to topic ( t_k )</td>
</tr>
<tr>
<td>( r_k = (r_{k1}, \ldots, r_{kK}) )</td>
<td>( r_{kk'} ), transition probability from topic ( t_k ) to topic ( t_{k'} )</td>
</tr>
<tr>
<td>( \alpha = (\alpha_1, \ldots, \alpha_K) )</td>
<td>parameters of the Dirichlet prior for the topic multinomial parameters ( \theta_1, \ldots, \theta_J )</td>
</tr>
<tr>
<td>( \beta = (\beta_1, \ldots, \beta_K) )</td>
<td>parameters of the Dirichlet prior for the word multinomial parameters ( \phi_1, \ldots, \phi_K )</td>
</tr>
<tr>
<td>( \gamma = (\gamma_1, \ldots, \gamma_K) )</td>
<td>parameters of the Dirichlet prior for topic transition multinomial parameters ( r_1, \ldots, r_K )</td>
</tr>
<tr>
<td>( \pi_{jw} = (\pi_{jw1}, \ldots, \pi_{jwK}) )</td>
<td>variational probabilities that a token of word ( v_w ) in document ( d_j ) is assigned to each topic</td>
</tr>
<tr>
<td>( \lambda_j = (\lambda_{j1}, \ldots, \lambda_{jK}) )</td>
<td>variational Dirichlet posterior parameters corresponding to ( \theta_j )</td>
</tr>
<tr>
<td>( \mu_k = (\mu_{k1}, \ldots, \mu_{kW}) )</td>
<td>variational Dirichlet posterior parameters corresponding to ( \phi_k )</td>
</tr>
<tr>
<td>( \nu_k = (\nu_{k1}, \ldots, \nu_{kK}) )</td>
<td>variational Dirichlet posterior parameters corresponding to ( r_k )</td>
</tr>
<tr>
<td>( \tau )</td>
<td>the mixing proportion of the topic multinomial parameters of the cited documents</td>
</tr>
<tr>
<td>( \omega_{j'j} )</td>
<td>the weight of document ( d_{j'} ) with respect to document ( d_j )</td>
</tr>
</tbody>
</table>

Existing topic models provide an analysis of reference relationships mainly with the following two approaches: establishing links among topic multinomial probabilities via Dirichlet processes [18, 24] or via logistic normal distributions[14, 23]. The inference for the former approach is complicated due to its nonparametric nature. This type of approach is not discussed any further in this paper. The latter approach is also intricate, because the model has a part of its parameters in \( K \times K \) dense covariance matrices, which need to be inverted. Therefore, covariance matrices used for logistic normal are often assumed to be diagonal [3, 1, 8]. One important reason this approach avoids dense matrices is that it considers topic transitions at many different places in observed data, e.g. at all citation links or at all pairs of adjacent time points along the time axis, as is discussed in Section 2.1. Consequently, many matrices should be taken into consideration in modeling, and thus the assumption of sparseness is required to accelerate inference.

In contrast, TERESA has a single \( K \times K \) matrix to model topic transition probabilities. Therefore, we have no reason to make this matrix sparse. However, our VB inference has a time complexity proportional to \( MK^2 \), where \( M \) is the number of citation links. Therefore, we have decided to use GPU to accelerate the inference.

### 3. PROPOSAL

#### 3.1 Generative Description

First, we describe how documents are generated by our topic model, TERESA. We extend LDA by introducing a corpus-wide topic transition matrix \( R \). The \((k, k')\)th entry \( r_{kk'} \) of \( R \) means the probability of the transition from topic \( t_k \) to topic \( t_{k'} \). TERESA generates documents as follows:

1. For each topic \( t_k \in T \), \( k = 1, \ldots, K \), draw a word multinomial distribution \( \text{Multi}(\phi_k) \) from the corpus-wide word Dirichlet prior distribution \( \text{Dir}(\beta) \). The \( w \)th entry \( \phi_{kw} \) of \( \phi_k \) is the probability that word \( v_w \in \mathcal{V} \) is used to express topic \( t_k \). The \( w \)th hyperparameter \( \beta_w \) of \( \text{Dir}(\beta) \) corresponds to word \( v_w \).

2. For each topic \( t_k \in T \), \( k = 1, \ldots, K \), draw a topic transition multinomial distribution \( \text{Multi}(r_k) \) from the corpus-wide topic transition Dirichlet prior distribution \( \text{Dir}(\gamma) \). The \( k' \)th entry \( r_{kk'} \) of \( R \) is the transition probability from topic \( t_k \) to topic \( t_{k'} \). The \( k' \)th hyperparameter \( \gamma_{k'} \) of \( \text{Dir}(\gamma) \) corresponds to topic \( t_{k'} \) appearing in the citing documents, not in the cited documents. The corpus-wide topic transition probability matrix \( R \) is obtained by setting the \( k \)th row of \( R \) to \( r_k' \), i.e., \( R = (r_1, \ldots, r_K)' \).

3. For each document \( d_j \in D \), \( j = 1, \ldots, J \),

   (a) Draw a topic multinomial distribution \( \text{Multi}(\theta_j) \) from the corpus-wide topic Dirichlet prior distribution \( \text{Dir}(\alpha) \). The \( j \)th entry \( \theta_{jk} \) of \( \theta_j \) is the probability that any word in document \( d_j \) is used to express topic \( t_k \). The \( k \)th hyperparameter \( \alpha_k \) of \( \text{Dir}(\alpha) \) corresponds to topic \( t_k \).

   (b) Let \( C_j \subseteq D \) be the set of the documents cited by document \( d_j \). We obtain a new topic multinomial distribution with parameters \( (\eta_{j1}, \ldots, \eta_{jK}) \) by combining topic multinomial parameters linearly as follows:

\[
\eta_j = (1 - \tau)\theta_j + \tau \sum_{j'} \omega_{jj'} R^{Tj'} \theta_{j'}
\]

where \( \omega_{jj'} \) is the weight of document \( d_{j'} \) with respect to document \( d_j \). \( \sum_{j'} \omega_{jj'} = 1 \) holds for every \( j \), where \( \omega_{jj'} = 0 \) for \( d_{j'} \notin C_j \). For simplicity, we set \( \omega_{jj'} = 1/C_j \) when \( d_{j'} \in C_j \). Obviously, TERESA is reduced to LDA when \( \tau = 0 \). In the evaluation experiment, we tested the values from 0.1 to 0.9 with step size 0.1 for \( \tau \).

(c) Let \( n_j \) be the length of document \( d_j \). For each word token \( x_{ji} \), \( i = 1, \ldots, n_j \), draw a topic from \( \text{Multi}(\eta_j) \) and set the drawn topic as the value of the latent variable \( z_{ji} \). Then, draw a word from \( \text{Multi}(\phi_{z_{ji}}) \) and set the drawn word as the value of the observed variable \( x_{ji} \).
3.2 Inference and Its Parallelization

Second, we describe how the posterior distributions of TERESA can be inferred. TERESA extends LDA by introducing a topic transition matrix $R$. However, this matrix $R$ is of size $O(K^2)$ and increases the computational burden of inference. While the time complexity of the inference for LDA is only proportional to $K$, that for TERESA is proportional to $K^2$. Therefore, we need to accelerate our inference.

We employ variational Bayesian inference (VB) [4] for an approximated estimation of posterior parameters, because VB is suitable for an efficient parallelization. Zhai et al. subtly discuss why VB can be efficiently parallelized [26]. However, we do not consider a parallelization with OpenMP and/or MPI, which is considered by Zhai et al. We consider a parallelization with Nvidia CUDA compatible devices, because our VB frequently performs the same operation on hundreds of different data simultaneously and favors SIMD architecture, which is implemented by Nvidia GPUs. In this paper, we assume that the input data can be stored on the device memory of GPU without being split into many subsets. Therefore, VB is not seriously affected by the latency during a data transfer between CPU and GPU, because such transfer occurs only before and after the computation on GPU, not in the course of the computation.

With respect to VB for LDA, the inaugural work by Blei et al. [4] contains detailed descriptions. Therefore, most details can be referred to [4]. Here we show the update formulae for the posterior parameters without derivation. Appendix A only gives derivations required for TERESA. Our approximated variational inference updates LDA posteriors and topic transition posteriors alternately.

Let $\pi_{jwk}$ be the variational posterior probability that word $v_w$ expresses topic $t_k$ in document $d_j$. Note that $\sum_k \pi_{jwk} = 1$ for every pair of $j$ and $w$. Let $\text{Dir}(\lambda_j)$ be the variational Dirichlet posterior distribution for the word multinomial distribution Multi($\theta_j$). Further, let $\text{Dir}(\mu_k)$ be the variational Dirichlet posterior distribution for the topic multinomial distribution Multi($\phi_k$). Then, based on the discussion in [4],

$$
\pi_{jwk} \propto \exp \left( \Psi(\lambda_{jk}) \right) \cdot \frac{\exp \left( \mu_{kw} \right)}{\exp \left( \mu_{k0} \right)}
$$

$$
\mu_{kw} = \beta_w + \sum_j n_{jw} \pi_{jwk}
$$

(2)

where $\Psi(\cdot)$ represents digamma function, $\mu_{k0}$ is defined to be $\sum_w \mu_{kw}$, and $n_{jw}$ is the number of the occurrences of word $v_w$ in document $d_j$. Obviously, $(\pi_{jw1}, \ldots, \pi_{jwK})$ can be updated in parallel for different pairs of $j$ and $w$. Further, $\mu_{kw}$ can be updated independently for different pairs of $k$ and $w$. Therefore, the updates in Eq. (2) are efficiently parallelized by using GPU.

With respect to $\lambda_{jk}$, we use the following update for LDA:

$$
\lambda_{jk} = \alpha_k + \langle n_{j,k} \rangle,
$$

(3)

where $\langle n_{j,k} \rangle = \sum_w n_{jw} \pi_{jwk}$. However, in TERESA, the generation of word tokens in document $d_j$ is affected not only by the topic distribution of $d_j$ but also by those of the documents cited by $d_j$. Therefore, we obtain a completely different update formula. Let $\text{Dir}(\nu_k)$ be the variational Dirichlet posterior for the topic transition multinomial Multi($r_k$). Then, $\lambda_{jk}$ can be updated by using the equation below:

$$
\lambda_{jk} \approx \alpha_k + (1 - \tau) \langle n_{j,k} \rangle + \frac{1}{\Psi'(\lambda_{jk})} \cdot \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'} \rangle \left\{ \Psi(\nu_{k'k}) - \Psi(\nu_{k0}) \right\}
$$

(4)

where $\nu_{k0} = \sum_k, \nu_{k'}$, $\lambda_{jk} = \sum_k \lambda_{jk}$, and $\Psi(\cdot)$ is trigamma function. Recall that $\omega_{j'j} = 0$ when document $d_j$ is not cited by document $d_{j'}$. Further, note that $\lambda_{jk}$ is equal to $\sum_k \alpha_k + n_j$ and thus does not depend on $\lambda_{jk}$. Eq. (4) is of the form $x = a + b/\Psi'(x)$, which can be solved by a
binary search for the constants $a$ and $b$. Therefore, based on Eq. (4), we can obtain an updated value for $\lambda_{jk}$. Since the right hand side of Eq. (4) does not contain $\lambda_{jk}$ for $k' \neq k$ and $\lambda_{0j}$ is a constant when $\alpha_k$s are fixed, we can update $\lambda_{jk}$ in parallel for different pairs of $j$ and $k$.

However, the approximation we introduce for obtaining Eq. (4) is not good when $\lambda_{jk}$ is small (cf. Eq. (14)). Consequently, $\lambda_{0j}$ becomes different from $\sum_j \alpha_k + n_j$ after updating each $\lambda_{jk}$. Therefore, we recalculate $\lambda_{j1}, \ldots, \lambda_{jK}$ to make $\lambda_{0j}$ equal to $\sum_j \alpha_k + n_j$ after an update of $K$ parameters $\lambda_{j1}, \ldots, \lambda_{jK}$ for a fixed $j$. Inferences with less approximation can be achieved by a gradient-based method. Although we implemented an optimization using L-BFGS [16, 10], the running time was prohibitively long, because we should call L-BFGS once for each document in each iteration to solve a $K$-dimensional optimization problem. Therefore, it is reserved as future work to make gradient-based inferences practical.

The topic transition posterior parameters $\nu_{kk'}$ are updated by using the following formula:

$$
\nu_{kk'} = \gamma_k + \tau \sum_j (n_{jk'} \sum_j' \omega_{j'} \lambda_{jk'}/\lambda_{0j})
$$

(5)

Since $\lambda_{jk}/\lambda_{0j}$ is the posterior probability of topic $t_k$ in document $d_{j'}$, $\nu_{kk'}$ is updated by using linear combinations of topic probabilities along citation links. This is occasionally similar to the proposal in [14]. The derivation of the above update formulae for $\lambda_{jk}$ and $\nu_{kk'}$ is included in Appendix A.

3.3 Implementation

We implemented the inference for TERESA as follows: 1) Initialize the assignment of each word token to a topic randomly selected from $\{t_1, \ldots, t_K\}$; 2) Run 500 iterations of collapsed Gibbs sampling (CGS) for LDA [6]; 3) Initialize the posteriors $\pi_{jwk}, \lambda_{jk}$, and $\mu_{k\omega}$ of TERESA by using the result of CGS, and initialize $\nu_{kk'}$ by using these initialized posteriors; and 4) Run 50 iterations of VB for TERESA. The Dirichlet hyperparameters are updated by Minka’s method [12].

We have tested many settings for GPU acceleration of VB. We obtained various wall-clock running times as in Table 2. Each running time is the time required for one iteration of VB, which is obtained by dividing the total running time of 50 iterations of VB by 50. We used GPU only for VB, not for the initialization by CGS for LDA. Table 2 also contains per-iteration running times obtained for Intel Core i7 CPUs. Therefore, we can compare the efficiency achieved by multithreading on Core i7 CPUs with that achieved by multithreading on GeForce GPUs. We fixed the number of GPU threads to 512 for ease of comparison between different types of GPU.

While it is known that collapsed variational Bayesian inference (CVB) [21, 2] is also suitable for parallelization by GPUs [25], it is reserved as future work.

4. EXPERIMENT

4.1 Data Sets

We employed two data sets available from the Web in our experiment. Their specifications are presented in Table 3.

The one data set is the CORA data set¹, often used for the experiments related to citation analysis. The other is the HEP-PH data set, a set of tex documents along with a citation graph available at KDD Cup 2003 Web site². Each tex document in the HEP-PH data set corresponds to a paper in the hep-ph portion of the arXiv. For both data sets, we removed stop words and applied a Porter’s stemmer. We did not remove any tex commands from the HEP-PH data set, because we wanted to know what kind of special symbols (e.g. \angle, \widehat, \dag, etc) are likely to be used in a particular discipline of physics.

We checked the soundness of inference by calculating perplexity [4] for 10% randomly chosen words, whose number is denoted by $N_{test}$. The perplexity is defined as

$$
\text{perplexity} = \exp \left\{ \sum_j \sum_i \log \sum_k \lambda_{jk} \mu_{k\omega_j} / N_{test} \right\}.
$$

(6)

We had a perplexity around 570 for the CORA data set after CGS for LDA, and the perplexity was not significantly modified by VB for TERESA. For the HEP-PH data set, we had a perplexity around 710 after CGS for LDA, and the perplexity was also not significantly modified by VB for TERESA. These are the perplexities obtained when $K = 300$. Since TERESA did not significantly modify the perplexity obtained as a result of CGS for LDA, we could extract a corpus-wide view of topic transitions by TERESA without affecting the generalization power of LDA.

4.2 Evaluation Method

As a quantitative measure for evaluating the quality of topic evolution analysis, we devise a new measure called diversity plus focusedness (D+F) based on the posterior probabilities of topic transitions. Note that the posterior probability of the transition from topic $t_k$ to topic $t_{k'}$ is $\nu_{kk'}/\nu_{0k}$. D+F score is defined by combining the two measures, transition diversity and expected focusedness:

- **Transition diversity.** Let $P(t_k) = \sum_j \lambda_{jk}/\sum_j \lambda_{jk}$ and $P_{Tr}(t_k|t_k) = \nu_{kk'}/\nu_{0k}$. $P(t_k)$ is the posterior probability of the occurrence of topic $t_k$. Then the posterior probability of the transition to topic $t_k$ can be written as $P_{Tr}(t_k) = \sum_{k'} P(t_{k'}) P_{Tr}(t_k|t_{k'})$. By using $P_{Tr}(t_k)$, we define transition diversity as follows:

$$
T_{Div} = - \sum_k P_{Tr}(t_k) \log P_{Tr}(t_k).
$$

(7)

<table>
<thead>
<tr>
<th>Table 2: Running Time per VB Iteration (sec).</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORA data set ($K = 300$)</td>
</tr>
<tr>
<td>8 threads on i7-2600K @ 3.40GHz</td>
</tr>
<tr>
<td>8 threads on i7 950 @ 3.07GHz</td>
</tr>
<tr>
<td>8 threads on i7 X 990 @ 3.47GHz</td>
</tr>
<tr>
<td>12 threads on i7-3930K @ 3.20GHz</td>
</tr>
<tr>
<td>1 block of size 512 on GTX 580 @ 1.54GHz</td>
</tr>
<tr>
<td>1 block of size 512 on GTX 570 @ 1.56GHz</td>
</tr>
<tr>
<td>HEP-PH data set ($K = 300$)</td>
</tr>
<tr>
<td>8 threads on i7 970 @ 3.20GHz</td>
</tr>
<tr>
<td>1 block of size 512 on GTX 580 @ 1.54GHz</td>
</tr>
<tr>
<td>1 block of size 512 on GTX 580 @ 1.59GHz</td>
</tr>
</tbody>
</table>

¹http://people.cs.umass.edu/~mccallum/data.html

²http://www.cs.cornell.edu/projects/kddcup/datasets.html
Figure 3: Two trimmed portions of topic evolution networks extracted by TERESA from the CORA data set for $K = 300$. Each topic is represented by the most frequent 21 words. Thicker arrows correspond to topic transitions of larger probability. VB was run on Nvidia GTX580 @ 1.59GHz. The running time of the inference containing CGS for LDA was around 9,200 seconds with Intel Core i7-2600K @ 3.40GHz.

$TrDiv$ is an entropy measure. When $TrDiv$ is large, the transition to each topic occurs equally often. This means that the topic model under evaluation can extract topics so that every topic joins a corpus-wide topic evolution equally. In other words, no topics are neglected in topic evolution analysis.

We define a measure called focussedness as follows:

$$Foc(t_k) \equiv \sum_{k'} P_{Tr}(t_k | t_{k'}) \log P_{Tr}(t_k | t_{k'})$$  \hspace{1cm} (8)

which is a negative entropy measure and is thus less than or equal to 0. When $Foc(t_k)$ is close to 0, only a limited number of topics are frequently reached from topic $t_k$. Further, we obtain an expected focussedness as follows:

$$EFoc \equiv \sum_k P(t_k) Foc(t_k)$$  \hspace{1cm} (9)

When $EFoc$ is large, the probability distribution of the transitions from any topic is highly skewed. In other words, only a limited number of transition edges starting from each node have a large probability.

$D+F$ is defined to be the sum of $TrDiv$ and $EFoc$. When $D+F$ is large, all topics are equally visited, and, at the same time, the choice of transition destination is highly selective. We contend that topic evolution analyses that can give larger $D+F$ are better. We give a rationale for this contention by considering three extreme cases.

First, assume that the transition probability matrix is the identity matrix. Then $TrDiv = -\sum_k P(t_k) \log P(t_k)$ and $EFoc = 0$. Therefore, $D+F$ is equal to the entropy $-\sum_k P(t_k) \log P(t_k)$, which is larger when $P(t_k)$s show less differences. In this case, we extract the topics that are totally “independent” in the sense that each topic only transits to itself. Of course, articles in reality may cite articles from heterogeneous research fields. However, it is desirable to extract relatively independent components where the transitions among different components rarely occur. Therefore, it can be said that a larger $D+F$ is better.

Second, assume that the transition probability matrix is a matrix all of whose columns except one are zero vectors. Then $TrDiv = 0$ and $EFoc = 0$. Therefore, $D+F = 0$. In our experiment, we found that smaller values of $D+F$ corresponded to an inference of poor quality. To be precise, it was observed that, when the inference was not successful, a large number of transitions pointed to a limited number of topics, which correspond to a vague content. Typically, such topics were represented by the words expressing vague concepts or by the words similar to stop words. In this case, the transition matrix had large values only in a limited number of columns. Therefore, it can be said that $D+F$ should be substantially larger than 0.

Third, assume that all entries of the transition probability matrix are $1/K$. Then $TrDiv = \log K$ and $EFoc = -\log K$. Therefore, $D+F$ is 0. This matrix means that all transitions are equally frequent, and thus the matrix corresponds to no meaningful topic evolution analyses. It is confirmed again that $D+F$ should be substantially larger than 0.

The latter two types of topic transition probability matrix is not informative, and the first one is, in a sense, an unreachable ideal. Therefore, we consider that a larger $D+F$ value...
corresponds to a better topic evolution analysis, though we cannot reach an extreme case where the transition matrix is the identity matrix. Each document set may have its own largest among the reachable D+F scores.

4.3 Evaluation Results

Figures 5 and 6 present the evaluation results for the CORA data set when \( K = 300 \) and \( K = 200 \), respectively, where we use the proposed measure D+F. Since these two figures make no large difference, we only provide the results for \( K = 300 \) with respect to the HEP-PH data set in Figure 7. All line graphs contain the results when we choose the value of \( \tau \) from 0.1 to 0.9 with step size 0.1. The horizontal axis represents \( \tau \). The vertical axis represents D+F score. We executed ten instances of VB inference, and the mean of the corresponding ten D+F scores is presented at each value of \( \tau \). When \( \tau = 0 \) and \( \tau = 1 \), D+F was 0, though not depicted in the graphs. The error bars on each graph show the width of plus/minus one standard deviation.

Solid and dashed lines show the results after 50 iterations and those after one iteration of VB inference, respectively. Therefore, dashed lines correspond to the cases where \( \nu_{kk'} \) is obtained directly from the results of CGS for LDA by applying Eq. (5). In contrast, solid lines correspond to the cases where we alternately update LDA posteriors and topic transition posteriors 50 times. Namely, the difference between the solid line and the dashed line in each graph corresponds to the difference between TERESA and LDA. As is shown in each graph, TERESA can give a better topic evolution analysis for smaller values of \( \tau \). However, when \( \tau \) is close to 1, TERESA seems to put too much emphasis on citation data and thus fails to achieve a balanced inference over text data and citation data. We can recommend 0.1 or 0.2 for the value of \( \tau \).

4.4 Examples of Topic Evolutions

We have already given an example of the results of topic evolution analysis in Figure 1, where we can find several interesting transitions. For example, the topic seemingly corresponding to machine translation (represented by translate, languag, system, english, machin, cross, automat, etc) and the topic seemingly corresponding to speech recognition (represented by speech, recognit, system, word, speaker, model, recogn, etc) point to each other with a thick arrow. That is, we can reveal that machine translation has a close bi-directional relationship with speech recognition based on the result given by TERESA.

More examples are presented in Figures 3 and 4. Figure 3 contains the examples of topic evolutions extracted from the CORA data set. On the top panel of Figure 3, the word network appears in several boxes. However, the boxes representing the topics related to network protocols (i.e., the boxes containing the words network, protocol, internet, servic, multicast, etc) and the boxes representing the topics related to neural networks (i.e., the boxes containing the words network, neural, learn, train, etc) are never connected. This conforms to our intuition. On the bottom panel of Figure 3, the topic seemingly related to Bayesian statistical analysis (represented by model, estim, data, distribut, statist, bayesian, paramet, etc) has incoming arcs both from the topic seemingly related to MCMC inferences (represented by chain, markov, carlo, mont, converg, etc), and from the topic seemingly related to bioinformatics (represented by sequence, protein, dna, molecular, align, etc). We can guess that some papers on Bayesian methods were written based on the preceding works related to MCMC inferences and also on the works related to bioinformatics.

We present the examples of topic evolutions extracted from the HEP-PH data set in Figure 4. On the left panel, the topic at the top left corner has the word dag as the most frequent word, which is the mathematical symbol “\( \dag \)” often used in expressing the unitarity of matrices. This topic and the topics that have incoming arcs from it are seemingly related to penguin diagram (represented by penguin, decai, ckm, asymmetri, etc). On the right panel, we can find a similar topic (represented by penguin, decai, asymmetri, etc) at the left center position. This topic has a thin arc coming from the topic represented by decai, quark, hadron, semilepton, etc, and this topic has bidirectional relationships with the topics that are also represented by semilepton, hadron, etc. However, the word semilepton does not appear on the left panel of Figure 4 at the position close to the topic related to penguin diagram. Therefore, we can guess that the papers discussing penguin diagram cite the papers on semileptonic decay only in a limited frequency, and thus that semileptonic decay is not a main topic of the papers cited by the papers on penguin diagram. In this manner, we can com-
pare topic evolution networks given by different executions of VB on the same data set to obtain a more robust and comprehensive observation.

4.5 Usage
As in case of LDA, TERESA gives different inference results for different random initialization. Figure 2 presents three different topic transition networks extracted from the HEP-PH data set by TERESA under completely the same setting, i.e., $K = 300$ and $\tau = 0.2$. However, TERESA can provide a corpus-wide view on how topics interrelate. Therefore, users of our method can run the inference several times on the same data set under the same setting and inspect the obtained topic evolution networks to crop interesting portions from each result and then to compare those cropped portions. In contrast, when we use the methods that give a pile of local topic transition patterns, as are enumerated in Section 2, this type of usage is hardly realizable, because it is difficult, in the first place, to integrate local views for obtaining a unified view on topic evolutions.

5. CONCLUSION
This paper provides an LDA-like topic model for extracting a corpus-wide view on how latent topics interrelate in a given set of scientific articles. While many existing approaches provide an accumulation of a large number of local views, our approach, TERESA, can give a single global view. Since topic modeling is primarily a method of dimensionality reduction, TERESA gives a “projection” of complicated semantical relationships among documents as a single transition matrix. Practically, we may run our method repeatedly on the same data set and obtain multiple “projections” to find interesting topic evolutions, as we have done in Section 4.4 with respect to the figures presented in this paper.

The only weak point of our method is the approximation introduced to obtain a non-gradient-based inference (cf. Eq. (14)). Therefore, the most important future work is to devise an inference with less approximation.

6. REFERENCES
Table 3: Specifications of Data Sets

<table>
<thead>
<tr>
<th>data set</th>
<th># docs</th>
<th># words</th>
<th># unique doc-word pairs</th>
<th># references</th>
<th># word tokens for training (for test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORA</td>
<td>36,183</td>
<td>8,542</td>
<td>1,523,768</td>
<td>129,492</td>
<td>2,127,005 (235,566)</td>
</tr>
<tr>
<td>HEP-PH</td>
<td>34,546</td>
<td>86,964</td>
<td>21,399,454</td>
<td>416,146</td>
<td>90,796,243 (10,084,293)</td>
</tr>
</tbody>
</table>


APPENDIX

A. DERIVATION

We have the full joint distribution of TERESA as follows:

\[
p(x, z, \theta, \phi, R | \alpha, \beta, \gamma) = \prod_j p(\theta_j | \alpha) \cdot \prod_k p(\phi_k | \beta) \times \prod_k p(r_k | \gamma) \cdot \prod_i \prod_j p(z_{ji} | \theta, R) p(x_{ji} | \phi_{z_{ji}})
\]

where we omit free parameters \( \tau \) and \( \omega_{ji} \). We apply a standard procedure of variational approximation as is provided in [4] by assuming that the variational posterior is fully factorized. The lower bound of the log evidence can be obtained by applying Jensen’s inequality as follows:

\[
\log p(x | \alpha, \beta, \gamma) \geq \int_z \sum_k q(z | \pi) q(\theta | \lambda) q(\phi | \mu) q(R | \nu) \\
\times \log p(x | z, \phi)p(z | \theta, R)p(\theta | \alpha)p(\phi | \beta)p(R | \gamma) q(z | \pi) q(\theta | \lambda) q(\phi | \mu) q(R | \nu) \\
\times d\theta d\phi d\nu (11)
\]

We denote the left hand side of Eq. (11) simply by \( L \). For convenience, we use the following notations: \( \lambda_{ji} \equiv \sum_k \lambda_{jk} \), \( \mu_{ij} \equiv \sum_k \mu_{jk} \), \( \nu_{0k} \equiv \sum_k \nu_{kk} \), \( \nu_{0j} \equiv \sum_k \pi_{jk} r_{ij} \), and...
Therefore, our approximation works for larger values of \( n_j \). We introduce an approximation to Eq. (13) by assuming

\[
\langle j | k \rangle \equiv \sum_w \pi_{jw} n_{jw}.
\]

Then \( \mathcal{L} \) can be expanded as follows:

\[
\begin{align*}
\mathcal{L} &= \sum_{j,k} \langle j | k \rangle \left\{ \Psi(\mu_{jk}) - \Psi(\mu_{ko}) \right\} \\
&+ (1 - \tau) \sum_{j,k} \langle j | k \rangle \left\{ \Psi(\lambda_{jk}) - \Psi(\lambda_{ko}) \right\} \\
&+ \sum_{j,k} \langle j | k \rangle \sum_{j'} \omega_{j'j} \lambda_{j'k} \left\{ \Psi(\nu_{j'k}) - \Psi(\nu_{k0}) \right\} \\
&+ J \log \Gamma(\alpha) - J \sum_{k} \log \Gamma(\alpha_k) - \sum_{j} \log \Gamma(\lambda_{j0}) \\
&+ \sum_{j,k} \log \Gamma(\lambda_{jk}) + \sum_{j,k} (\alpha_k - \lambda_{jk}) \left\{ \Psi(\lambda_{jk}) - \Psi(\lambda_{j0}) \right\} \\
&+ K \log \Gamma(\beta_0) - K \sum_{w} \log \Gamma(\beta_w) - \sum_{j} \log \Gamma(\mu_{ko}) \\
&+ \sum_{j,k} \log \Gamma(\mu_{jk}) + \sum_{k,w} (\beta_w - \mu_{jk}) \left\{ \Psi(\mu_{jk}) - \Psi(\mu_{ko}) \right\} \\
&+ K \log \Gamma(\gamma_0) - K \sum_{k',k} \log \Gamma(\gamma_{k'}) - \sum_{k} \log \Gamma(\nu_{k0}) \\
&+ \sum_{k,k'} \log \Gamma(\nu_{kk'}) + \sum_{k,k'} (\gamma_{k'} - \nu_{kk'}) \left\{ \Psi(\nu_{kk'}) - \Psi(\nu_{k0}) \right\} \\
&- \sum_{j,w} \sum_{k} \pi_{jwk} \log \pi_{jwk},
\end{align*}
\]

where the third term is obtained by an additional application of Jensen’s inequality to the log of the linear combination of topic multinomial parameters in Eq. (1).

It is a complicated process to optimize \( \mathcal{L} \) with respect to \( \lambda_{jk} \). By differentiating \( \mathcal{L} \) with respect to \( \lambda_{jk} \), we obtain:

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial \lambda_{jk}} &= \{ (1 - \tau) n_j - (\lambda_{jk} - \alpha_k) \} \Psi'(\lambda_{jk}) \\
&- \{ (1 - \tau) n_j - (\lambda_{j0} - \alpha_0) \} \Psi'(\lambda_{j0}) \\
&+ \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'} \rangle \{ \Psi'(\nu_{j'k'}) - \Psi'(\nu_{k0}) \} \\
&- \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'} \rangle \sum_{k} \lambda_{jk} \{ \Psi'(\nu_{kk'}) - \Psi'(\nu_{k0}) \}.
\end{align*}
\]

We introduce an approximation to Eq. (13) by assuming

\[
\lambda_{jk} \Psi'(\lambda_{jk}) \approx \lambda_{j0} \Psi'(\lambda_{j0})
\]

and \( \lambda_{j0} \). Consequently, we obtain the following derivative:

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial \lambda_{jk}} &\approx \{ (1 - \tau) n_j - (\lambda_{jk} - \alpha_k) \} \Psi'(\lambda_{jk}) \\
&- \{ (1 - \tau) n_j - (\lambda_{j0} - \alpha_0) \} \Psi'(\lambda_{j0}) \\
&+ \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'} \rangle \frac{\Psi'(\nu_{kk'}) - \Psi'(\nu_{k0})}{\Psi'(\lambda_{jk})} \\
&- \frac{\tau}{\lambda_{j0}} \sum_{j'} \omega_{j'j} \sum_{k'} \langle n_{j'k'} \rangle \sum_{k} \frac{\Psi'(\nu_{kk'}) - \Psi'(\nu_{k0})}{\Psi'(\lambda_{jk})}.
\end{align*}
\]

Eq. (15) tells that \( \partial \mathcal{L}/\partial \lambda_{jk} \) is almost equal to 1 for large values of \( \lambda_{jk} \).

It is an important future work to propose an inference avoiding the approximation we introduce here, because we apply this approximation as a last resort to achieve a non-gradient-based and thus efficient inference. Although a gradient-based inference was once implemented, the running time was unacceptably long, because \( K \) dimensional optimization problem should be solved for each of the \( J \) documents.

Further, we differentiate \( \mathcal{L} \) with respect to \( \nu_{kk'} \) and obtain:

\[
\begin{align*}
\frac{\partial \mathcal{L}}{\partial \nu_{kk'}} &= \tau \Psi'(\nu_{kk'}) \sum_{j} \langle n_{jkk'} \rangle \sum_{j'} \omega_{j'j} \lambda_{j'k} \frac{\lambda_{jk}}{\lambda_{j'0}} \\
&- \tau \Psi'(\nu_{k0}) \sum_{k'} \sum_{j} \langle n_{jkk'} \rangle \sum_{j'} \omega_{j'j} \lambda_{j'k} \frac{\lambda_{jk}}{\lambda_{j'0}} \\
&+ (\gamma_{kk'} - \nu_{kk'}) \Psi'(\nu_{kk'}) - \Psi'(\nu_{k0}) \sum_{k} (\gamma_{kk'} - \nu_{kk'}).
\end{align*}
\]

Therefore, \( \nu_{kk'} \) can be updated as:

\[

nu_{kk'} = \gamma_{kk'} + \tau \sum_{j} \langle n_{jkk'} \rangle \sum_{j'} \omega_{j'j} \lambda_{j'k} \frac{\lambda_{jk}}{\lambda_{j'0}}
\]

Figure 8 is a plot of the function \( f(x) = x \Psi'(x) \). As this graph shows, \( f(x) \) is almost equal to 1 for large values of \( x \). Therefore, our approximation works for larger values of \( \lambda_{jk} \),