On Modelling Non-linear Topical Dependencies

Zhixing Li ADAM0730@GMAIL.COM
Siqiang Wen WENSQ2329@GMAIL.COM
Juanzi Li LIJUANZI@TSINGHUA.EDU.CN
Peng Zhang ZPJUMPER@GMAIL.COM
Jie Tang JIETANG@TSINGHUA.EDU.CN
Department of Computer Science, Tsinghua University, Beijing, China

Abstract

Probabilistic topic models such as Latent Dirichlet Allocation (LDA) discover latent topics from large corpora by exploiting words' co-occurring relation. By observing the topical similarity between words, we find that some other relations, such as semantic or syntax relation between words, lead to strong dependence between their topics. In this paper, sentences are represented as dependency trees and a Global Topic Random Field (GTRF) is presented to model the non-linear dependencies between words. To infer our model, a new global factor is defined over all edges and the normalization factor of GRF is proven to be a constant. As a result, no independent assumption is needed when inferring our model. Based on it, we develop an efficient expectation-maximization (EM) procedure for parameter estimation. Experimental results on four data sets show that GTRF achieves much lower perplexity than LDA and linear dependency topic models and produces better topic coherence.

1. Introduction

Latent Dirichlet Allocation (LDA), first proposed by Blei et al. in 2003 (Blei et al., 2003), is one of the most widely used probabilistic topic models. In the past ten years, it has been successfully used to analyze document collections, images (Chi-Chun & Prasenjit, 2011), music (Hu & Saul, 2009) (Shalit et al., 2013) and videos (Weinshall et al., 2013).

As pointed out by Blei (Blei, 2012), LDA makes several assumptions. One unrealistic assumption is that the words in a document are "exchangeable". It implies that, given a prior topical mixture, the topics of words in a document are conditionally independent. Many extensions have been proposed to relax this assumption but most of them are limited to linear topical dependencies between words. Gruber (Gruber et al., 2007) assumes that the topic prior of the words in a sentence is dependent on its proceeding sentence's. Zhu (Zhu & Eric, 2010) assumes that topic assignment of a words is dependent on its neighboring words with similar syntax features. However, words may be dependent with each other in a much more complex manner. In text processing, words may depend on each other in a tree structure according to linguistic knowledge(Sartorio et al., 2013). In image processing, superpixels are related with each other spatially(Li & Li, 2007).

To confirm the existence of topical dependencies in texts, we analyzed the documents of Reuters-21578 using standard LDA (topic number = 10) and then conducted a statistic on the similarity between words on topics. The results are illustrated in Figure 1.

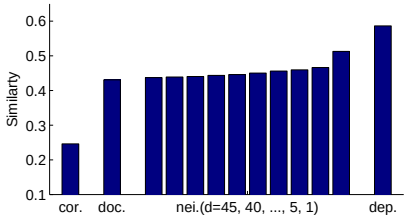


Figure 1. Topical similarity between words using LDA

Figure 1 shows the average topical similarity between words in the same corpus (cor.), document (doc.), neighbring area (nei.) and depended word pairs (dep.) respectively. We have following observations. First, the topical similarity between words in the same document is much higher than that in the same corpus. It proves that positive correlation exists among the topics of words in the same document and it is the basis of LDA. Second, the topi-

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

cal similarity between neighboring words increases within a decreasing distance ranging from 45 to 1. It implies that neighboring relation is more significant than co-occurring relation. This has been proven by many existing literatures(Zhu & Eric, 2010)(Gruber et al., 2007). Third, the topics of semantically or syntactically dependent word pairs achieve the highest similarity. It means that the syntactic or semantic dependencies lead to the strongest topical dependencies and they are able to provide more useful information in topic modeling.

However, modeling these topical dependencies is a non-trivial task because of the non-linear structures. The main challenge is the generation probability of topic assignments. The nature of Bayesian model requires that the topic assignments should be drawn from a legal probability measure to make sure that the model won't lead to unexpected biases. However, for distributions that can model the probability of a set of (mutually) dependent variables, such as Markov Random Field (MRF) and Conditional Random Field (CRF), the calculation of the normalization factor is usually difficult and to some extent impossible in some scenarios. To overcome this challenge, CTRF (Zhu & Eric, 2010) makes an extra independent assumption between word pairs. Even though, it can only model linear topical dependencies between words. Graber (Boyd-Graber & Blei, 2008) proposes a model to make use of the syntactic information available from parse trees but it only handles unidirectional dependencies.

In this paper, we present a novel method to model non-linear structure topical dependencies. The key of our model is the embedding of Global Random Field (GRF) for the sampling of latent topics over words. Except the multinomial factors used in conventional LDA model, GRF imports a new global factor defined over all edges to model the topical dependencies between words. With some constrains that can be easily satisfied, the normalization factor of GRF is proven to be a constant. As a result, no independent assumption is needed when inferring our model. We develop an EM algorithm for parameter estimation and experimental results on four different copra show that non-linear dependencies do improve the modelling performance when comparing with existing methods.

The rest of this paper is organized as follows. Section 2 gives a brief review of related works and Section 3 presents our model. Inference and parameter estimation are presented in Section 4 and in Section 5 we conduct experiments on four corpora. Section 6 concludes this paper.

2. Related Work

In this section we give a brief review of related works. Table 2 defines some frequently occurring variables.

Symbol	Description
w:	a word, or a vertex in a graph G
z:	a topic, or a state of vertex w.
θ:	the topic mixture of a document.
d:	a document that is composed of a sequence of words.
N:	the number of words in a document.
K:	the number of topics.

Table 1. Notations of some frequently occurring variables.

2.1. LDA and its extensions

To our knowledge, although there are lots of studies focusing on modeling the topical dependencies for LDA, there is few existing work that models topical dependencies in graph structure. Most of them are of linear chain structure or unidirectional tree structure.

LDA (Blei et al., 2003) is a generative three-layer Bayesian model assuming that the topics of words in the same document are conditionally independent. In LDA, the probability of a topic sequence is:

$$p_{lda}(\mathbf{z}|\theta) = \prod^{n} Multi(z_n|\theta) \tag{1}$$

HTMM (Gruber et al., 2007) model assumes that all words in the same sentence should be assigned to the same topic and the topic of one sentence is dependent on its preceding sentence. For a document, it generates a topic sequence using a Markov process:

$$p_{lda}(\mathbf{z}|\theta) = p(z_1|\theta) \prod_{2}^{n} p(z_n|z_{n-1}, \theta) \tag{2}$$

The strategy of seqLDA (Lan et al., 2010) and STM (Boyd-Graber & Blei, 2008) is similar but in seqLDA, the topical dependencies are defined on longer units such as chapters and in STM, words are generated conditioned on their parents in the parse trees.

CTRF (Zhu & Eric, 2010) relaxes the independent assumption by defining a linear chain Conditional Random Field on the topic sequence. It differs from HTMM and STM that the topical dependencies in CTRF are mutual instead of unidirectional. CTRF defines the probability of a topic sequence using Generalized Linear Model (GLM):

$$p_{ctrf}(\mathbf{z}|\theta, a) = \frac{\prod_n [\phi(z_n|\theta, a)\phi(z_n, z_{n+1}|\theta, a)]}{\sum_{\mathbf{z}'} \prod_n [\phi(z'_n|\theta, a)\phi(z'_n, z'_{n+1}|\theta, a)]} \tag{3}$$

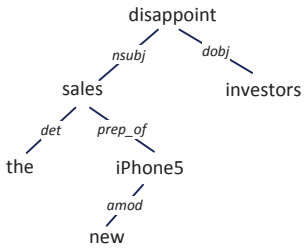
However, the complexity of the log-partition function $Z(\theta) = \sum_{\mathbf{z}'} \prod_n [\phi(z'_n|\theta, a)\phi(z'_n, z'_{n+1}|\theta, a)]$ prevents accurate inference of CTRF. CTRF makes an assumption that the potential functions $\phi(\cdot)$ of different words (pairs) are independent with each other to facilitate the calculation of the expectation of $Z(\theta)$.

There are some lines of research aiming at relaxing other assumptions of LDA. Newman (David et al., 2011) proposed a model using the relation of words directly to modify the distribution of word generation instead of topical dependencies. Griffiths (Griffiths et al., 2005) presents a model that automatically identifies if a word is a functional word or a content word using a Hidden Markov Model (HMM). Markov Topic Random Field (MTRF) (Hal, 2009) assums that documents are dependent with each other. Posterior Regulization (Ganchev et al., 2010) defines constrains on the latent variables to reduce the searching space. Our model differs from these methods by modeling mutual topical dependencies between words in non-linear structure.

2.2. Graphic Representation of Documents

In most existing research, a document is represented as an unordered collection of words, which is also known as "bag of words" representation. In these studies, the relations between words are totally abandoned except the co-occurring relation. Usually, this representation is used to simplify the the modeling while according to linguistic knowledge, words are dependent with each other in a more complex manner.

Dependency parsing (Ryan & Joakim, 2011) is a natural language processing tool which can find the syntax or semantic relations between words, and then represents a sentence into a tree (See Figure 2 for example). Most of the linked words in this tree are tightly related with each other, e.g., *sales* and *IPhone5*, *disappoint* and *invertors*. In practice, we can filter out the edges that link to function words such as *the* and add new edges using some rules. For example, the *subject* and *object* of the same verb. In this way, with the help of a dependency parser, a sentence can be transformed into a graph conveniently.



Sentence: *The sales of new iPhone5 disappoint investors.*

Figure 2. A dependency tree

The graphic representation of documents reserves more information than "bagof words" representaion and actually has imported linguistic knowledge in it. In the next section, we propose a new model based on a carefully designed random field to make use of the non-linear semantic or syntactic dependencies between words in topic model.

3. Global Topic Random Field

In this section, we present Global Topic Random Field (GTRF), a new model that can exploit topical dependency of arbitrary structures. Given a document d, we at first transform it into a graph G_d using the method discussed in Section 2.2.

3.1. Global Random Field

Once a document is represented as a graph, the sampling of topics can be cast into the sampling of this graph. We propose a random field called Global Random Field (GRF) to model the sampling process. Before presenting GRF, we start by defining a new distribution.

Theorem 1: Given an undirected graph $G =< W, E >$ where $W = \{w_i | i = 1, 2, ..., n\}$ is a set of vertices, $E = \{(w'_i, w''_i) | i = 1, 2, ..., m\}$ is the edge set and the state of a vertex w is drawn from a finite set $Z = \{z_i | i = 1, 2, ..., k\}$, function:

$$
\begin{aligned}
P(G) = f_G(g) = \frac{1}{|E|} \prod_{w \in W} \phi(z_w) \times \sum_{(w', w'') \in E} \phi(z_{w'}, z_{w''}) \\
s.t.\ 1.\ \phi(z) > 0,\ \phi(z, z') > 0 \\
2.\ \sum_{z \in Z} \phi(z) = 1 \\
3.\ \sum_{z', z'' \in Z} \phi(z')\phi(z'')\phi(z', z'') = 1
\end{aligned}
\tag{4}
$$

is a probability measure.

In Equation 4, $\phi(\cdot)$ is a function defined on a single vertex and $\phi(\cdot, \cdot)$ is defined on an edge. g is one sample (topic assignment) of G and z_w is the state (topic) of vertex w.

Proof: Let's consider a graph $G' =< W, E' >$ that contains all vertices but only one edge in G. Without loss of generality, we let $E' = \{(w_1, w_2)\}$ and $W^{\neg} = W - \{w_1, w_2\}$, then we have:

$$
\begin{aligned}
f_{G'}(g) &= \prod_{w \in W} \phi(z_w) \times \phi(z_{w_1}, z_{w_2}) \\
&= \prod_{w \in W^{\neg}} \phi(z_w) \times [\phi(z_{w_1})\phi(z_{w_2})\phi(z_{w_1}, z_{w_2})]
\end{aligned}
\tag{5}
$$

Summing $f_{G'}$ over all possible g, we obtain:

$$
\begin{aligned}
&\sum_g f_{G'}(g) = \sum_g \left[\prod_{w \in W^{\neg}} \phi(z_w) \times [\phi(z_{w_1})\phi(z_{w_2})\phi(z_{w_1}, z_{w_2})] \right] \\
&= \prod_{w \in W^{\neg}} \sum_{z_w \in Z} \phi(z_w) \times \sum_{z_{w_1}, z_{w_2} \in Z} \left[\phi(z_{w_1})\phi(z_{w_2})\phi(z_{w_1}, z_{w_2}) \right] \\
&= 1
\end{aligned}
\tag{6}
$$

By summing over all G' (one for an edge):

$$\sum_{g} f_G(g) = \frac{1}{|E|} \sum_{g} \sum_{G'} f_{G'}(g) = 1 \tag{7}$$

Obsviouly, $f_G(g) > 0$ and therefore it is a legal probability measure.

We call a random field with a distribution as in Equation 4 a *Global Random Field* (GRF) because the item $\sum_{(w',w'')\in E} \phi(z_{w'}, z_{w''})$ sums over through all edges in G while in MRF or CRF, each factor (potential function) is defined on a clique. One may find that there is no normalization factor in Equation 4. This is the advantage of GRF with which in the modeling process, we can avoid the complex calculation of normalization factor and thus do not need to make extra independent assumptions.

Since there is no constrain on the structure of G, all kinds of structures are acceptable, ranging from simple structure such as linear chain to complex structures such as tree or network.

3.2. Modeling topical dependencies Using GRF

Based on GRF, we propose a new model *Global Topic Random Field* (GTRF). GTRF differs from standard LDA and its extensions in the generation of words' topics.

In GTRF, for document d, given a topic mixture θ (that drawn from a Dirichlet prior) and its graphic representation $G_d = \{V, E\}$, the probability of the topic sequence $\boldsymbol{z}$ of d is modeled by:

$$\begin{aligned} p_{gtrf}(\boldsymbol{z}|\theta) = \frac{1}{|E|} \prod_{w\in V} Multi(z_w|\theta) \\ \times \sum_{(w',w'')\in E} \left(\sigma_{z_{w'}=z_{w''}}\lambda_1 + \sigma_{z_{w'}\neq z_{w''}}\lambda_2\right) \\ where\ \sigma_x = \begin{cases} 1 & x \text{ is } true \\ 0 & x \text{ is } false \end{cases} \end{aligned} \tag{8}$$

Comparing with Equation 4, we can see that the function defined on the topic of a single word is a Multinomial distribution parameterized by θ and the function defined on a edge is $(\sigma_{z_e^1=z_e^2}\lambda_1 + \sigma_{z_e^1\neq z_e^2}\lambda_2)$. Obviously, $\sum_z Multi(z|\theta) = 1$ and $Multi(z|\theta) > 0$. To satisfy the third constraint, the following equation should hold true:

$$\begin{aligned} &\sum_{z_{w'}, z_{w''}} \big[Multi(z_{w'}|\theta) Multi(z_{w''}|\theta) \\ &\qquad \times (\sigma_{z_{w'}=z_{w''}}\lambda_1 + \sigma_{z_{w'}\neq z_{w''/'}}\lambda_2)\big] \\ &= \sum_{z_{w'}=z_{w''}} \big(Multi(z_{w'}|\theta) \times Multi(z_{w''}|\theta) \times \lambda_1\big) \\ &\qquad + \sum_{z_{w'}\neq z_{w''}} \big(Multi(z_{w'}|\theta) \times Multi(z_{w''}|\theta) \times \lambda_2\big)\big) \\ &= \theta^T\theta\lambda_1 + (1-\theta^T\theta)\lambda_2 = 1 \end{aligned} \tag{9}$$

To make Equation 9 true, we have:

$$\lambda_1 = \lambda_2 + \frac{1-\lambda_2}{\theta^T\theta} \tag{10}$$

According to Theorem 1, once Equation 10 is satisfied, the function defined in Equation 8 is a legal probability measure.

Given a topic sequence $\boldsymbol{z}$, we can divide the edge set E of G_d into two sets: E_C and E_{NC}. E_C contains edges that connect two vertices that have been assigned the same topic and E_{NC} contains the remains. The edges in E_C are called **Coherent Edges**. Then Equation 8 can be rewritten as:

$$\begin{aligned} p_{gtrf}(\boldsymbol{z}|\theta) &= \frac{1}{|E|} \prod_{w\in V} Multi(z_w|\theta) \times (|E_C|\lambda_1 + |E_{NC}|\lambda_2) \\ &= \prod_{w\in V} Multi(z_w|\theta) \times \left(\frac{|E_C|(1-\lambda_2)}{|E|\theta^T\theta} + \lambda_2\right) \end{aligned} \tag{11}$$

To derive Equation 11, we used Equation 10 and the fact $|E| = |E_C| + |E_{NC}|$.

According to our observation, the linked words have higher topical similarity. To model positive correlations between topics of linked words, **coherent edges should be rewarded**. This can be satisfied by choosing a $\lambda_2 < 1$ and lower λ_2 means higher reward to coherent edges.

Given distribution P_{gtrf}, then we have the generation procedure of a document as follows:

0. Transform document d into a graph G_d.
1. Draw $\theta \sim Dir(\alpha)$.
2. Draw a topic sequence $\boldsymbol{z} \sim P_{gtrf}(\boldsymbol{z}|\theta)$
3. For each of N word w_n in d:
   draw $w_n \sim Multi(\beta_{z_n,w_n})$.

In this generation procedure, the topics for a document are sampled at the same time. Topical dependencies between related words are modeled using GRF which will reward graphs that contain more coherent edges.

4. Inference and Estimation

In this section we discuss how to infer the posterior distribution and estimate parameters of GTRF. Although in the conditional probability $p_{gtrf}(\boldsymbol{z}|\theta)$, there is no log-partition function and as a result the summation over an exponential number of latent topic assignments can been avoided, $p_{gtrf}(\boldsymbol{z}|\theta)$ contains a global factor summing over all the edges which is difficult to calculate. This factor is approximated using Taylor series in this paper.

4.1. Posterior Inference

Like LDA and its extensions, GTRF can not be inferred exactly. We develop a variational inference method for GTRF. At first we give the probability of a document:

$$p(\boldsymbol{w}|\alpha,\beta) = \int p(\theta|\alpha)\sum_{\mathbf{z}} p_{grtf}(\mathbf{z}|\theta)\prod_{n} p(w_n|z_{w_n},\beta)\mathrm{d}\theta$$

Parameters α and β can not be estimated directly due to the coupling between θ and β. We develop a variational distribution q to approximate p:

$$q(\theta,\mathbf{z}|\gamma,\varphi) = Dir(\theta|\gamma)\times\prod_{n} Multi(z_{w_n}|\varphi_{w_n})$$

Following the deduction of standard LDA in (Blei et al., 2003), we can write the likelihood of a document in GTRF into:

$$\begin{aligned} L \triangleq\ & \mathrm{E}_q\big[\log p(\theta|\alpha)\big] + \mathrm{E}_q\big[\log p(\mathbf{z}|\theta)\big] + \mathrm{E}_q\big[\log p(\boldsymbol{w}|\mathbf{z},\beta)\big] \\ & - \mathrm{E}_q\big[\log q(\theta)\big] - \mathrm{E}_q\big[\log q(\mathbf{z})\big] \end{aligned}$$

All items except the second one can be expanded as in LDA. Here we only expand the second item to save space:

$$\begin{aligned} & \mathrm{E}_q\big[\log p(\mathbf{z}|\theta)\big] \\ =\ & \mathrm{E}_q\left[\log\Big(\prod_n Multi(z_{w_n}|\theta)\times\frac{|E_C|(1-\lambda_2)+\lambda_2|E|\theta^T\theta}{|E|\theta^T\theta}\Big)\right] \\ =\ & \mathrm{E}_q\big[\log\big(\prod_n Multi(z_{w_n}|\theta)\big)\big] \\ & + \mathrm{E}_q\big[\log\big(|E_C|(1-\lambda_2)+\lambda_2|E|\theta^T\theta\big)\big] \\ & - \mathrm{E}_q\big[\log\big(|E|\theta^T\theta\big)\big] \end{aligned} \tag{12}$$

In the last line, the first item is the same as the likelihood in p_{lda} w.r.t the variational distribution. The second and third item can not be calculated directly so we use Taylor series to approximate them and we have:

$$\begin{aligned} & \mathrm{E}_q\big[\log\big((1-\lambda_2)|E_C|+\lambda_2|E|\theta^T\theta\big)\big] - \mathrm{E}_q\big[\log\big(|E|\theta^T\theta\big)\big] \\ \approx\ & \mathrm{E}_q\big[\zeta_1^{-1}\big((1-\lambda_2)|E_C|+\lambda_2|E|\theta^T\theta\big) - 1 + \log\zeta_1\big] \\ & - \mathrm{E}_q\big[\zeta_2^{-1}\big(|E|\theta^T\theta\big) - 1 + \log\zeta_2\big] \\ =\ & \left(\frac{1-\lambda_2}{\zeta_1}\right)\mathrm{E}_q(|E_C|) - \left(\frac{\zeta_1-\zeta_2\lambda_2}{\zeta_1\zeta_2}|E|\right)\mathrm{E}_q(\theta^T\theta) \\ & + \log\zeta_1 - \log\zeta_2 \end{aligned} \tag{13}$$

In Equation 13, we use the fact that $|E|$, which is the number of edges in G_d, is a constant w.r.t G_d. The reward given to coherent edges is controled by λ_2.

$\mathrm{E}_q|E_C|$ can be calcualted as:

$$\mathrm{E}_q|E_C| = \sum_{(w_n,w_m)\in E}\varphi_{w_n}^T\varphi_{w_m} \tag{14}$$

$\mathrm{E}_q(\theta^T\theta)$ can be obtained according to the definition of Dirichlet distribution:

$$\mathrm{E}_q(\theta^T\theta) = \sum_{K}^{K}\mathrm{E}_q\theta_i^2 = \sum_{K}^{K}\frac{\gamma_i(\gamma_i+1)}{\gamma_0(\gamma_0+1)} \tag{15}$$

where $\gamma_0 = \sum_i^K \gamma_i$.

The rest items in L can be deducted in the same way as LDA.

4.2. Parameter Estimation

In the previous subsection, we have completed the deduction of the likelihood L and it can be represented as a function of $(\gamma,\varphi,\alpha,\beta,\zeta_1,\zeta_2,\lambda_2)$. Among these parameters, ζ_1,ζ_2 are used for Taylor approximation and they can be embedded with the values of $(1-\lambda_2)|E_C|+\lambda_2|E|\theta^T\theta$ and $|E|\theta^T\theta$ in the previous iteration respectively. λ_2 is an hyper parameter and it's value is determined by data observations or other methods. The updating rules for α and β are the same as in LDA and we omit them to save space. Here we give the updating rules for φ.

$$\varphi_{w_n i}\propto\beta_{iv}\exp\left(\Psi(\gamma_i)+\frac{1-\lambda_2}{\zeta_1}\times\sum_{(w_n,w_m)\in E}\varphi_{w_m i}\right) \tag{16}$$

In Equation 16, i is the index of topic and v is the index of w_n in the vocabulary and w_m is one of the words that semantically or syntactically dependent on w_n. From this equation, we can see that once $\lambda_2<1$, φ_{w_n}, the topic distribution of word w_n will obtain positive mass from w_m and therefore after iterations the topic distributions of connected words will approach the same.

Unfortunately, we can not obtain the direct updating rule for γ. As a suboptimal strategy, we update γ using Newton method and here is the derivation of L w.r.t. γ:

$$\begin{aligned} \frac{\partial L}{\partial\gamma_i} =\ & \big(\alpha_i-\gamma_i+\sum_n\phi_{w_n i}\big)\big(\Psi'(\gamma_i)-\Psi'(\gamma_0)\big) \\ & - \frac{\zeta_1-\zeta_2\lambda_2}{\zeta_1\zeta_2}|E|\left(\sum_K\frac{\gamma_i(\gamma_i+1)}{\gamma_0(\gamma_0+1)}\right)' \end{aligned} \tag{17}$$

where $\gamma_0 = \sum_i^K\gamma_i$.

Clearly, the time complexity of $\frac{\partial L}{\partial\gamma_i}$ is $O(N+K)$ where N is the number of words of current document and K is the

number of topics. Therefore, $\frac{\partial L}{\partial \gamma_i}$ can be compute efficiently.

With the aboving updating rules, we leverage an EM algorithm to estimate parameters α and β and the procedure is substantially the same as in LDA. At the very beginning, α and β are averagely sampled. Then for each iteration, in E-step, the algorithms find the best γ and φ for current α and β; in M-step, α and β are updated using the obtained γ and φ.

5. Experiments

In this section, we conduct several experiments to compare GTRF model with standard LDA and CTRF. We choose LDA and CTRF for two reasons. First, LDA is chosen as a baseline that makes no use of relation between words. Second, as far as we know, CTRF is the only exiting model can model mutual topical dependencies between words which is similar to our model. STM, HTMM and SeqLDA are nice models but they focuses unidirectional dependencies between words or chapter so they are not chosen as comparison models in this paper.

5.1. Datasets

We use four datasets in our experiments, two are news documents and two are research papers. The four datasets used in this paper are:

Reuters-21578[1]: It contains 21,578 documents appeared on Reuters newswire in 1987.

20NewsGroups[2]: A collection of approximately 20,000 newsgroup documents.

NIPS data[3] (A. et al., 2007): The accepted papers of NIPS from 2000 to 2005.

ICML data: The accepted papers of ICML from 2007 to 2013.

Table 5.1 illustrates these four datasets. As illustrated, these four datasets are of two different kinds. Retures-21578 and 20NewsGroups contain more documents but with shorter length while NIPS and ICML data contain less documents but with longer length. What's more important, the former two datasets are news articles so they may cover more topics than the two research paper datasets. We choose these two kinds of datasets intentionally to test GTRF's performance in different scenarios.

To implement GTRF, we parse all documents using Stanford Parser[4] (Marneffe et al., 2006). New edges are added between the *subject* and *object* of the same verb and stopwords and the edges connecting to them are removed after dependency parsing.

Dataset	# of docs	# of unique words	Avg # of words per doc
Reuters-21578	16,844	13,054	75
20NewsGroup	18,828	27,456	108
NIPS	1,000	15,415	1,704
ICML	1,004	20,907	2,599

Table 2. Illustration of datasets

5.2. Experiments setup

Following most existing studies on topic model, we test our model and comparison methods in document modeling and evaluate their performance using **predicative perplexity** (Blei et al., 2003). For all datasets, we train models with two thirds of documents and calculate predicative perplexity on the unseen one third of documents.

Because the topics of ICML and NIPS data are more concentrated and the number of documents is fewer, it is not proper to assume that they contains large number of topics. Therefore, we test all three models on ICML and NIPS data with topic numbers $K = 10, 15, 20, 25$. For the other two dataset, we test all three models with topic number $K = 10, 20, 50, 100$. In our GTRF model, there is a control parameter λ_2 that can not be estimated directly and we test GTRF with $\lambda_2 = 0.2, 0.4, 0.6, 0.8$.

5.3. Experimental results and analysis

We compare GTRF with existing models in three ways. First, we compare their performance on document modeling with perplexity. Then we test the topical similarity of words modelled by GTRF to find if GTRF can better model the dependencies between words in the documents. At last we conduct a case study to show the topical coherence of our model.

Document Modeling

Figure 3 and Figure 4 illustrates the experimental results. We repeat each experiment five times and the perplexity values shown here are the average values. Unsurprisingly, both CTRF and GTRF perform much better than standard LDA in all four corpora because both CTRF and GTRF uses extra information rather than co-occurring relation to discover topics.

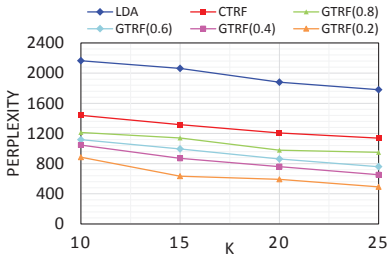
The comparison between the results of CTRF and GTRF is interesting. For the ICML and NIPS data, GTRF produces

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

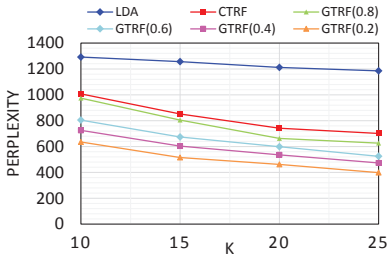
[2] http://qwone.com/ jason/20Newsgroups/

[3] http://ai.stanford.edu/ gal/Data/NIPS/

[4] http://nlp.stanford.edu/software/lex-parser.shtml

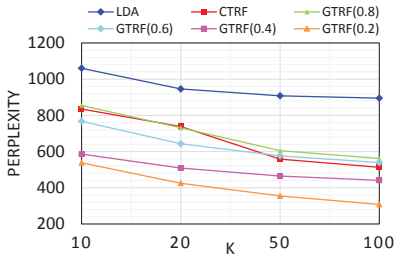


(a) *ICML*

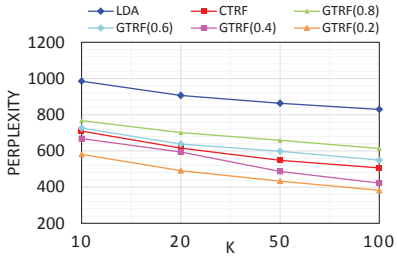


(b) *NIPS*

Figure 3. Experimental results on ICML and NIPS data.



(a) *20NewsGroup*



(b) *Reuters-21578*

Figure 4. Experimental results on 20NewsGroup and Reuters-21578.

much lower predictive perplexity than CTRF when λ_2 values from 0.8 to 0.2 (see Figure 3). While for Reuters-21578 and 20NewsGroups data, CTRF performs better than GTRF when λ_2 is greater (>0.6)(see Figure 4). As discussed in Section 3, lower λ_2 means higher reward to coherent edges, so one possible reason is that the length of news articles is short so that dependency between related words is not so significant comparing with neighboring words. In fact, we can see that the improvements of CTRF and GTRF on ICML and NIPS data are greater than on Reuters-21578 and 20NewsGroups. It can be explained with the same reason.

According to the results illustrated in Figure 3 and 4, we can see that lower λ_2 always leads to lower perplexity. However, it doesn't mean that λ_2 should be given a tiny value (e.g., 0.1 or even smaller). In experiments, we find that GTRF is to some extent sensitive to λ_2 and lower λ_2 tends to lead to lower stability. Table 3 shows the variation coefficient of the perplexity for different λ_2 (averaged on all 4 datasets). The variation coefficient of perplexities is calculated by:

$$vc(perp) = \frac{\sigma(perp)}{\mu(perp)} \tag{18}$$

where $\sigma(\cdot)$ is the standard derivation and $\mu(\cdot)$ is the mean. The coefficient of variation is a normalized measure (w.r.t the mean). Higher variation coefficient indicates more dispersion exists from the average value, which means lower stability to our method.

λ_2	0.2	0.4	0.6	0.8
vc	0.2287	0.1422	0.0946	0.0756

Table 3. The variation coefficient of λ_2.

From Table 3 we can see that the variation coefficient of the perplexity increases with the decreasing of λ_2. This suggests that although lower λ_2 produces lower perplexity, the stability of GTRF decreases as well. The most possible reason is that we reward all coherent edges evenly while it shouldn't be. In our future work we will try to classify edges according to lexical or statistic features and then reward them in different manners.

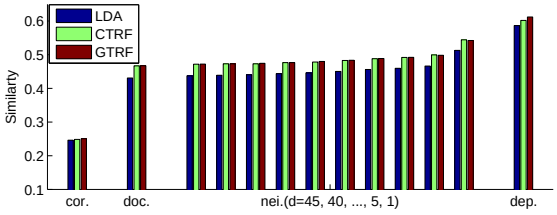


Figure 5. Topical similarity between words using GTRF.

Topical Similarity

As an echo of the data observation illustrated in Figure 1, Figure 5 shows the topic similarity between words modeled by GTRF, CTRF and LDA on Reuters-21578 (topic number = 10). Comparing with what shown in Figure 1, the topical similarity in the same corpus keeps almost

The U.S. congress and the oil industry are deeply divided on ways the government should assist the industry, hurt by the sharp fall in oil prices, and the subsequent growth in oil imports, industry analysts said. "The industry is deeply divided between those who support an oil tariff and those who believe tax incentives are better." Said Daniel Yergin, director of Cambridge Energy research associates, which recently completed a survey of the U.S. congress on energy issues.

The U.S. congress and the oil industry are deeply divided on ways the government should assist the industry, hurt by the sharp fall in oil prices, and the subsequent growth in oil imports, industry analysts said. "The industry is deeply divided between those who support an oil tariff and those who believe tax incentives are better." Said Daniel Yergin, director of Cambridge Energy research associates, which recently completed a survey of the U.S. congress on energy issues.

Figure 6. An example of the comparison between GTRF and LDA.

(top: LDA, bottom:GTRF. Each color stands for a topic.)

the same while the similarity in GTRF and CTRF between co-occurring words, neighboring words are much higher. In particular, GTRF outperforms CTRF on the similarity between dependent words. No one would deny that these word pairs should have higher topical similarity. This contrast proves that GTRF complies human cognition better than CTRF and LDA.

Topical Coherence

Figure 6 shows the comparison of topic assignments of GTRF ($K = 10, \lambda_2 = 0.4$) and LDA ($K = 10$) on the same document. In LDA, a word is assigned to the same topic no matter how many times it occurs and no matter what its context is. As a contrast, GTRF can assign different topics to different occurrences according to the context. For example, in the first row, GTRF assigns "oil" and "industry" the same topic and thus leads to higher topic coherent. We also test the result of CTRF, it can to some extent assign same topic to adjacent words but failed to deal with long distance word pairs or triples such as "government ... assist... industry".

In conclusion, the reason for the good performance of GTRF is on two-fold. First, GTRF integrates more information than existing models. GTRF is to some extent a semi-supervised model because the dependency parser is trained on human annotated data. Therefore, importing dependency parser is somehow equivalent to importing human linguistic knowledge. Secondly, when inferring GTRF, we make no extra assumption which means GTRF will treat different priors evenly. However, there is one more thing need further research: the control parameter λ_2. Although in experiments we observed that a smaller λ_2 will reward coherent edges more and leads to lower perplexity, it will brings potential risk to our model. In a more realistic manner, we should reward coherent edges discriminatively according to their lexical or statistical features instead of reward them evenly. This is our future work.

6. Conclusions

In this paper, we have proposed a novel model, Global Topic Random Field (GTRF) that aims at discoverring latent topics from large achieve of documents by exploiting topical dependencies between semantically or syntactically dependent words. While existing models assume that words in the same document are evenly related, GTRF makes a more reasonable assumption that these words should have higher topical similarity. We had investigated data observations to confirm this assumption. To model the complex dependency structures that can not be modeled by existing models, we have proposed GRF, a random field with a simple probability function. By integrating GRF into LDA, our GTRF model can both model complex structures and be inferred conveniently. We also have developed a variational inference and efficient EM algorithm to estimate GTRF's parameters and conducted series of experiments on different kinds of corpora. Experimental results show that GTRF achieves significantly better performance than existing start-of-the-art models.

Acknowledgment

The work is supported by 973 basic program research (No. 2014CB340504), NSFC (No. 61035004, No. 61222212, No. 61073073), NSFC-ANR (No. 61261130588), XLike (FP7-288342), and THU-NUS NExT Co-Lab.

References

A., Globerson, G., Chechik, F., Pereira, and N., Tishby. Euclidean Embedding of Co-occurrence Data. *The Journal of Machine Learning Research*, 8:2265–2295, 2007.

Blei, David, Ng., Andrew, and Jordan, Michael. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

Blei, David M. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April 2012.

Boyd-Graber, Jordan and Blei, D. Syntactic topic models. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, 2008.

Chi-Chun, Pan and Prasenjit, Mitra. Event detection with spatial latent dirichlet allocation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pp. 349–358, 2011.

David, Newman, V., Bonilla Edwin, and L., Buntine Wray. Improving topic coherence with regularized topic models. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, pp. 496–504, 2011.

Ganchev, Kuzman, Graça, João, Gillenwater, Jennifer, and Taskar, Ben. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11:2001–2049, August 2010.

Griffiths, Thomas L., Steyvers, Mark, Blei, David M., and Tenenbaum, Joshua B. Integrating topics and syntax. In *In Proceedings of the 17th Advances in Neural Information Processing Systems*, pp. 537–544, 2005.

Gruber, Amit, Weiss, Yair, and Rosen-Zvi, Michal. Hidden topic markov models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistic*, 2007.

Hal, Daumé III. Markov random topic fields. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 293–296, 2009.

Hu, Diane and Saul, Lawrence K. A probabilistic topic model for unsupervised learning of musical key-profiles. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pp. 441–446, 2009.

Lan, Du, Lindsay, Buntine Wray, and Huidong, Jin. Sequential latent dirichlet allocation: Discover underlying topic structures within a document. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 148–157, 2010.

Li, Cao and Li, Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of IEEE Internatinoal Conference in Computer Vision*, 2007.

Marneffe, Marie-Catherine De, Maccartney, Bill, and Manning, Christopher D. Generating typed dependency parses from phrase structure parses. In *In Proceedings of International Conference on Language Resources and Evaluation*, pp. 449–454, 2006.

Ryan, McDonald and Joakim, Nivre. Analyzing and integrating dependency parsers. *Computational Linguistic*, 37(1):197–230, March 2011.

Sartorio, Francesco, Satta, Giorgio, and Nivre, Joakim. A transition-based dependency parser using a dynamic parsing strategy. In *Proceeding of the 51st Annual Meeting of the Association for Computational Linguistics*, August 2013.

Shalit, Uri, Weinshall, Daphna, and Chechik, Gal. Modeling musical influence with topic models. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 244–252, May 2013.

Weinshall, Daphna, Levi, Gal, and Hanukaev, Dmitri. Lda topic model with soft assignment of descriptors to words. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 711–719, May 2013.

Zhu, Jun and Eric, Xing. Conditional topic random fields. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.