Relational Topic Model for Congressional Bills Corpus

You Lu
Department of Computer Science
University of Colorado Boulder
you.lu@colorado.edu

Shudong Hao
Department of Computer Science
University of Colorado Boulder
shudong.hao@colorado.edu

Abstract

In this project, we study the topic model with document networks, and implement the algorithm proposed called relational topic model. Additionally, we create a document network obtained from interactive user study, which is more reasonable than the ones used in the original paper, and is more appropriate for the assumption behind relational topic model. By analyzing the network using relational topic model, we validated the results presented in the user study.

1 Introduction

Documents networks (Croft et al., 1983), such as citation networks of documents, and hyperlinked networks of web pages, are becoming more and more prevalent in modern machine learning applications. The statistical analysis of these networks can provide both useful predictive models and descriptive statistics.

Traditional network analysis methods (Kemp et al., 2004; Hofman and Wiggins, 2008; Airoldi et al., 2008) only focus on the network properties, i.e., the link structure, of the network. Even though powerful, these methods ignore the properties of the nodes, which are as important as the link structure. For example, in a document citation network, the links are the citation relationships, and the attributes of a node is the words in that document. These two properties of a network are of the same importance.

Probabilistic topic models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003a) and hierarchical Dirichlet processes (Teh et al., 2012) and so forth, are popular generative models for analyzing data sets such as documents, images and videos. Traditional topic models only model the document content but do not take into account the connections between them.

To this end, Jonathan Chang developed relational topic model (Chang and Blei, 2009), which is a variant of LDA and can incorporate the links information into the model to improve the topic properties as well as predict links based on observed words. Relational topic model sees each edge as undirected and each link in the network as a binary random variable. However, in their original paper, they conducted experiments on Cora data (McCallum et al., 2000) and WebKB (Craven et al., 1998), which are two directed networks. As we will discuss later in section 5, these are not appropriate datasets, because the model assumes undirected links.

To better test the relational topic model, in this project, we create an undirected document network. We use Congressional Bills (Adler and Wilkerson, 2006) as our corpus. User labels generated by ALTO (Poursabzi-Sangdeh et al., 2016) as the edges of the network. Detailed information of our network will be discussed in the fifth section. Our second work in this project is to implement the relational topic model in Python. Finally, we use relational topic model to fit our new document network, so that we can validate the published results.

The paper is organized as follows. In the second section, we introduce the contributions of our projects. In the third and fourth sections, we briefly review the corresponding backgrounds. The fifth section illustrated our empirical results. Some discussions are in the sixth section. Researches most related to our project will be discussed in the seventh section. The last section is the conclusion of this project.

2 Contributions

Our project's contributions are three-fold.

2.1 Implemented the Model

First, we implemented relational topic model algorithm. We note that this model has more variation versions recently (Gui et al., 2014; Guo et al., 2015; Chen et al., 2015), but we still implement the basic version of the algorithm, proposed in Chang and Blei (2009). Gibbs sampling (Hrycej, 1990) and variational inference (Beal, 2003) are both popular methods for inferecing document-topic distributions θ and topic-word distributions ϕ.

Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm (Szymanski, 1987), commonly used in graphic models and Bayesian inference. It begins with ranom samples, and sample each observation based on conditional distributions of other observations. One drawback of Gibbs sampling is that is usually requires long time to converge to a steady state. When the graph has too many variables to sample and/or the dataset is too huge, the computational time required by Gibbs sampling is very expensive.

For relational topic model, however, the variational inference can be faster to converge. To deal with large dataset, it can be scaled easily as well. Therefore, in our project, we use variational inference to this end.

2.2 Created a Better Network

For typical research and study related to network structure and document networks, two typical datasets are *Cora* data and *WebKB* data. Cora data (McCallum et al., 2000) is obtained from Cora research search engine (the data set is available at: `https://hpi.de/naumann/projects/repeatability/datasets/cora-dataset.html`). Each document represents the abstract of a paper, and all of them build a citation network, where the documents are linked according to the citations in the reference of the paper. WebKB data (Craven et al., 1998) is a network built on webpages' hyperlink (the data set is available at: `http://www.cs.cmu.edu/~webkb/`). We treat each webpage (obtained from computer science departments websites) as a document, and use the hyperlink in those pages to build the network.

These networks are used in network studies as well as the original paper that proposed relational topic model. Apparently, these two networks are direct graphs, which means the links between documents are directive. For example, a hyperlink in a webpage might point out to another webpage, but the reverse is not necessarily true. In Cora data, the citation is also directed. We argue that, for relational topic model, direct graphs are not appropriate to use, because the assumption of relational topic model is based on undirect graphs. It is true that the relational topic model could be modified to consider direct graph situations, but for the basic model, it is reasonable to find a dataset that is consistent with the assumption.

Based on this idea, we create a new network from the user study in Poursabzi-Sangdeh et al. (2016). Different from the networks previously mentioned, this network is based on the content of the documents. Users annotate documents with labels, and when two documents are labeled by the same label, we connect the documents with a link. The construction of this network will be described in details in section 4.

2.3 Validated the Published Results

In the experiments in Poursabzi-Sangdeh et al. (2016), the authors use interactive topic modeling to speed document labeling. Then, by setting different number of topics across the whole corpus, they conclude that when the number of topics is 19, the performance is the best. Validating the result is relatively hard when the experiments involves human interaction. However, by using the network we constructed from the dataset they used and the annotations from the interaction, we are able to examine the work in a different perspective, that is, the network behind the interaction.

The project report is organized in the following way. First, we introduce the relational topic model with variational inference. Since it's based on the topic model, and more specifically, latent Dirichlet allocation (LDA), we also briefly review the generative process and algorithm of LDA. Then, we describe the construction and background of the network we used in our project in section 4. Next, we use

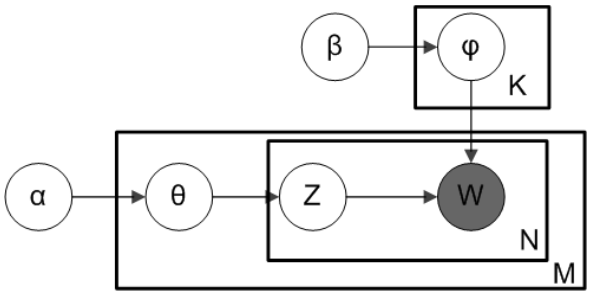

Figure 1: Graphical model of LDA. Source: Wikipedia.

the model to validate the results presented in Poursabzi-Sangdeh et al. (2016). In the end, we discuss our project in terms of evaluation, future work, and possible contribution to future research.

3 Relational Topic Model

In this section, we first briefly review the topic model used in relational topic model, latent Dirichlet allocation (LDA). Then we describe how to integrate topic model into network research by introducing relational topic model.

3.1 Latent Dirichlet Allocation

Topic model (Blei and Lafferty, 2006) such as Latent Semantic Indexing (LSI) (Letsche and Berry, 1997) and probabilistic LSI, uses word co-occurrence information across documents and the whole corpus to discover topics. LDA, proposed in Blei et al. (2003b), is the most popular one, which uses graphical model to model latent variables. Figure 1 shows the plate notation of LDA.

In LDA, we assume that each topic ϕ is a distribution over the vocabulary. When we interpret the topic, we look at the words of those with highest probabilities in the distribution, and use them to interpret the topic. For example, if a topic has high probabilities of words *cells*, *genes*, and *DNA*, we might interpret the topic as *biology*.

We also assume each document has a topic distribution θ. Thus, by looking at the topic distribution, we are able to infer what the main theme of the document is. For example, if the topic distribution of document d is heavily skewed to a topic that we interpreted as *music*, we might conclude that d is about *music* and *art*.

We give the generative process of the model in Algorithm 1.

Algorithm 1 Latent Dirichlet Allocation

1: **for** each document $\mathbf{d}_i \in \mathcal{D}$ **do**
2: Choose topic distribution $\theta_d \sim \text{Dir}(\alpha)$
3: **for** each topic $k \in \{1, 2, \ldots, K\}$ **do**
4: Choose topic-word $\phi_k \sim \text{Dir}(\beta)$
5: **for** each document $\mathbf{d}_i \in \mathcal{D}$ **do**
6: **for** each token $w_{d,i} \in \mathbf{d}_i$ **do**
7: Choose topic $k_{d,i} \sim \text{Mult}(\theta_d)$
8: Choose a word $w_{d,i} \sim \text{Mult}(\phi_{k_{d,i}})$

3.2 Relational Topic Model

Topic model provides a reasonable way to measure how "close" of documents are to each other. For example, a document talking about *global warming* should have similar document-topic distribution θ to another one with similar theme. This idea gives us an insight on document network study. In document citation networks, two documents are linked together is mostly because the content of the two documents are related to each other. Therefore, the topics of each document should be a reasonable way to construct a document network, and it is also a hint on discovering why some documents are linked by our observation.

Based on this idea, Chang and Blei (2009) proposed relational topic model that focuses on how topics across the corpus could help in document network study.

Intuitively, the documents with similar topic distributions should be linked together. We formalize this idea by adding an observed variable $y_{d,d'}$ to denote the link between document d and document d'. If they are linked together, $y_{d,d'} = 1$; if not, $y_{d,d'} = 0$. The graphical model is shown in Figure 2.

Following the generative process in the original LDA model, we also view the links are generated during the generating of documents. We show the generative process of relational topic model in Algorithm 2.

3.3 Link Probability Functions

In line 10 in Algorithm 2, we use $\psi(\cdot|z_d, z_{d'})$ to draw a link between documents, and we call this function **link probability function**. There are many options for this function, and

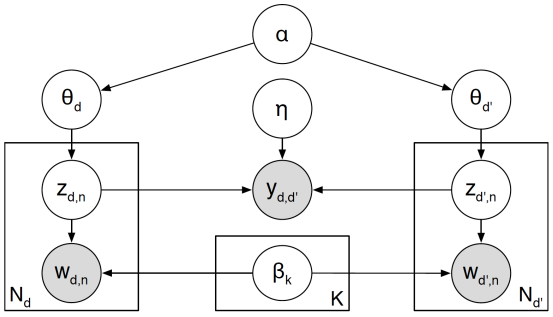

Figure 2: Plate notation of relational topic model. Source: Chang and Blei (2009).

Algorithm 2 Relational topic model

for each document $\mathbf{d}_i \in \mathcal{D}$ **do**
  Choose topic distribution $\theta_d \sim \text{Dir}(\alpha)$
for each topic $k \in \{1, 2, \ldots, K\}$ **do**
  Choose topic-word $\phi_k \sim \text{Dir}(\beta)$
for each document $\mathbf{d}_i \in \mathcal{D}$ **do**
  for each token $w_{d,i} \in \mathbf{d}_i$ **do**
    Choose topic $k_{d,i} \sim \text{Mult}(\theta_d)$
    Choose a word $w_{d,i} \sim \text{Mult}(\phi_{k_{d,i}})$
for each document pair d, d' **do**
  Draw a binary link $y \sim \psi(\cdot|z_d, z_{d'})$

different choices would give different performance. In the paper Chang and Blei (2009), two link probability functions are given.

The first one is based on sigmoid function:

$$\psi(y = 1|z_d, z_{d'}) = \sigma\left(\eta^\top(\overline{z_d} \circ \overline{z_{d'}}) + \nu\right) \quad (1)$$

$$\overline{z_d} = \frac{1}{N}\sum_{n=1}^{N} z_{d,n} \quad (2)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

where $\circ$ is the element wise product of two vectors. Intuitively, we can explain this link function in this way: first, we use $\overline{z_d} \circ \overline{z_{d'}}$ to measure how similar the two documents are by looking at their topic distributions. Then, we parameterize this measurement by η and ν. To case this result to a probability setting, we use sigmoid function $\sigma(\cdot)$ to project the results to the interval of $[0, 1]$.

The second one is similar to this setting. The only difference is that we replace the sigmoid function with exponential function:

$$\psi(y = 1|z_d, z_{d'}) = \exp\left(\eta^\top(\overline{z_d} \circ \overline{z_{d'}}) + \nu\right). \quad (4)$$

Note that the results returned by this function does not fit in the definition of probability, so we need some minor modification of the results to use it as a probability that two documents have a link. In this project, however, we implement the simplest method by using sigmoid function.

3.4 Variational Inference

In our project, we only implemented the RTM with exponential function. Hence, we only review the training method for it.

Given a corpus of D documents, the goal of training RTM is to compute its posterior:

$$\begin{aligned} & p(w, y, z, \theta|\beta, \alpha, \eta) \\ &= \prod_{d=1}^{D} p(\theta_d|\alpha)p(z_d|\theta_d)p(w_d|\beta, z_d) \\ &\times \prod_{d=1}^{D}\prod_{d'=1}^{D} p(y_{d,d'}|\eta, z_d, z_{d'}) \end{aligned} \quad (5)$$

However, this posterior is intractable. In the original paper, the authors uses a variational EM algorithm to approximate the posterior. The method first uses variational infrence (VI) to approximate the document specific posterior, and then updates the global parameters based on the local variational parameters. VI posits a factorized variational distribution:

$$q(z, \vartheta|\phi, \gamma) = \prod_{d=1}^{D} q(\theta_d|\gamma_d)\prod_{n=1}^{N_d} q(z_{dn}|\phi_{dn}) \quad (6)$$

and updates the variational parameters to minimize the KL divergence between the true local posterior and the variational distribution. The update rules for the local variational parameters are:

$$\begin{aligned} \phi_{dnk} &\propto \exp\{\sum_{d' \neq d} \frac{\eta \circ \bar{\phi}_{d'}}{N_d}\} \\ &\times exp\{\Psi(\gamma_{dk}) + log(\beta_{kw_{dn}})\} \end{aligned} \quad (7)$$

$$\gamma_{dk} = \alpha + \sum_{n=1}^{N_d} \phi_{dnk} \tag{8}$$

where $\Psi(.)$ is the digamma function and $\bar{\phi}_d$ represents the mean of ϕ_d. That is,

$$\bar{\phi}_d = \frac{1}{N_d} \sum_{n=1}^{N_d} \phi_{dn}$$

.

The update for global parameters in relational topic model is a little different from the traditional variational EM. Relational topic model only incorporates the observed links into the model. To do this, the authors add a regularization penalty parameter, that is, ρ, into the training. Hence, the update rules for the link function parameters are:

$$\begin{aligned} \upsilon &= \log(M - \mathbf{1}^{\boldsymbol{T}} \bar{\prod}) \\ &- \log(\rho(1 - \mathbf{1}^{\boldsymbol{T}} \bar{\pi}_\alpha) + M - \mathbf{1}^{\boldsymbol{T}} \bar{\prod}) \end{aligned} \tag{9}$$

$$\eta = \log(\bar{\prod}) - \log(\bar{\prod} + \rho \bar{\pi}_\alpha) - \mathbf{1}\upsilon \tag{10}$$

where

$$M = \sum_{(d,d')} 1,$$

$$\bar{\prod} = \sum_{(d,d')} \bar{\pi}_{d,d'},$$

$$\bar{\pi}_{d,d'} = \bar{\phi}_d \circ \bar{\phi}_{d'},$$

and

$$\bar{\pi}_\alpha = \frac{\alpha}{\mathbf{1}^{\boldsymbol{T}} \alpha} \circ \frac{\alpha}{\mathbf{1}^{\boldsymbol{T}} \alpha}.$$

The $\circ$ denotes the Hadamard (element-wise) product.

Finally, the update rule for the topic is:

$$\beta_{kv} \propto \sum_{d=1}^{N_d} n_v \phi_{dnv} \tag{11}$$

Note that here we use ϕ_{dv} instead of ϕ_{dn} in Equation 6, since the same term v have the same ϕ_{dn}.

4 User Annotated Network

As discussed before, one drawback of the original work in Chang and Blei (2009) is that the relational topic model does not consider direction of links in a network, but the datasets used are direct graphs. To make experiments more reasonable, we construct a network from the results in Poursabzi-Sangdeh et al. (2016). In this section, we describe how we construct this network.

4.1 Dataset

The dataset we use is from US congressional bills (available at `https://www.govtrack.us/`). The original dataset contains labels and sublabels, such as *health* and *agriculture* and so forth. These labels are used as gold standard answers in Poursabzi-Sangdeh et al. (2016). In order to construct a reasonable network **and** validate the user study results using relational topic model, we take the labels from users instead. By this, we are able to see if the network created by user is consistent with the gold standard labels, and thus confirm the results in Poursabzi-Sangdeh et al. (2016).

4.2 Annotating Documents

For each user, we let him/her to annotate a certain amount of documents during a 40-minute session. For each document, the user could:

- Annotate with an existing label. The existing labels might come from the user's previously created labels, or other users' labels.
- Annotate with a new label. In this setting, the user type into his/her own label, and it will be saved in the system.
- Skip the document.

We show the interface used for collecting annotations from users in Figure 3.

After each session, we collect a bunch of documents with annotations. When the whole experiment finishes, each document has been labeled by different users, and has its own label set. This label set is the critical point for constructing a network, and this will be introduced in the subsequent subsection.

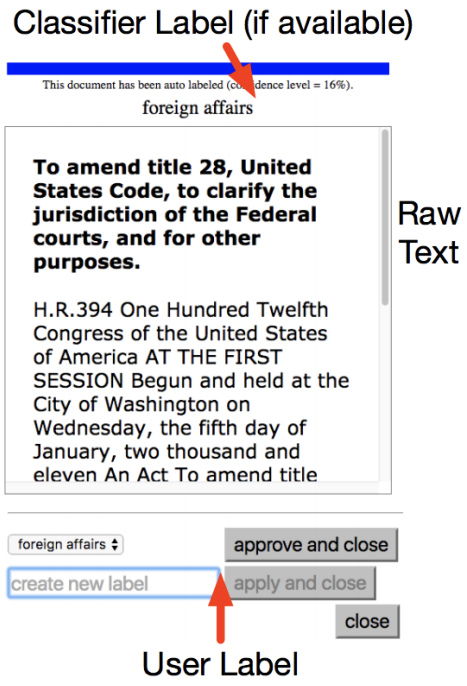

Figure 3: The interface for document annotation. Source: Poursabzi-Sangdeh et al. (2016).

4.3 From Annotations to Network

The labels from users are important hints of a document network. Although the concept of "document network" is not mentioned in the original paper, the understanding of those documents from crowdsourcing should reflect a network structure of the corpus: documents with similar themes are linked together.

Based on this idea, we use the annotations obtained from users to create a document network. In this network, as usual, we treat each document as a node v. When two documents have same labels, they are linked together, that is, an edge e is added between them.

As we will discuss in section 6.1, there are some drawbacks in this construction of network. However, this way of construction is still reasonable and effective.

5 Experiments

In this section, we introduce the experiment settings first. Then we present the results and our analysis on them.

5.1 Experiment Settings

In our data set, we have 40 user label sets. Each user uses his label set to annotate some documents. The total number of labeled documents is 1588. The total number of labels in these label sets is 1106. The statistical results of our data set is in Table 1. A subgraph of the Congressional Bills network is illustrated in Figure 4.

5.2 Evaluations

Following (Chang and Blei, 2010), we use the predictive label rank as our metric. Given the words of a held-out document, we compute the probability that it will link to each other document. We then rank the other documents according to this probability. The predictive rank is the average rank of the documents to which the held-out document actually did link. Lower rank is better.

$$PLR(d) = \frac{1}{D} \sum_{(d,d') \in E} r_{d'} \tag{12}$$

where E is the set of edges of the network. The $r_{d'}$ is the rank of the edge (d, d').

In all our experiments, we use ten fold cross-validation to assess the predictive label ranks. That is, we randomly seperate the data set into 10 subsets. At each time, we use nine subsets as the training set to train the relational topic model, and use the remaining one subset as the test set to compute predictive link rank. This process repeats 10 times, and we use the average predictive label ranks as the final reporting result.

5.3 Results

We show results in Figure 5 and Figure 6. First, we change the values of $alpha0$ from 1 to 5 at intervals of 1, and the number of topics K from 5 to 30 at intervals of 5. Recall that α is the prior for document-topic distribution, which means it encodes the prior belief of the distributions. In the figure, we use $alpha0$ to denote the sum of α for all the topics. That is, when $alpha0 = K \cdot \alpha$ where K is the number of topics. As we see in the figure, when $alpha0 = 1$ and $K = 20$, the predicted label rank is the lowest, which means in this network, when number of topics is 20, we get the best performance.

Next, we show the results of ρ in Figure 6. Recall that ρ is used as regularization penalty, in order to consider the negative observations

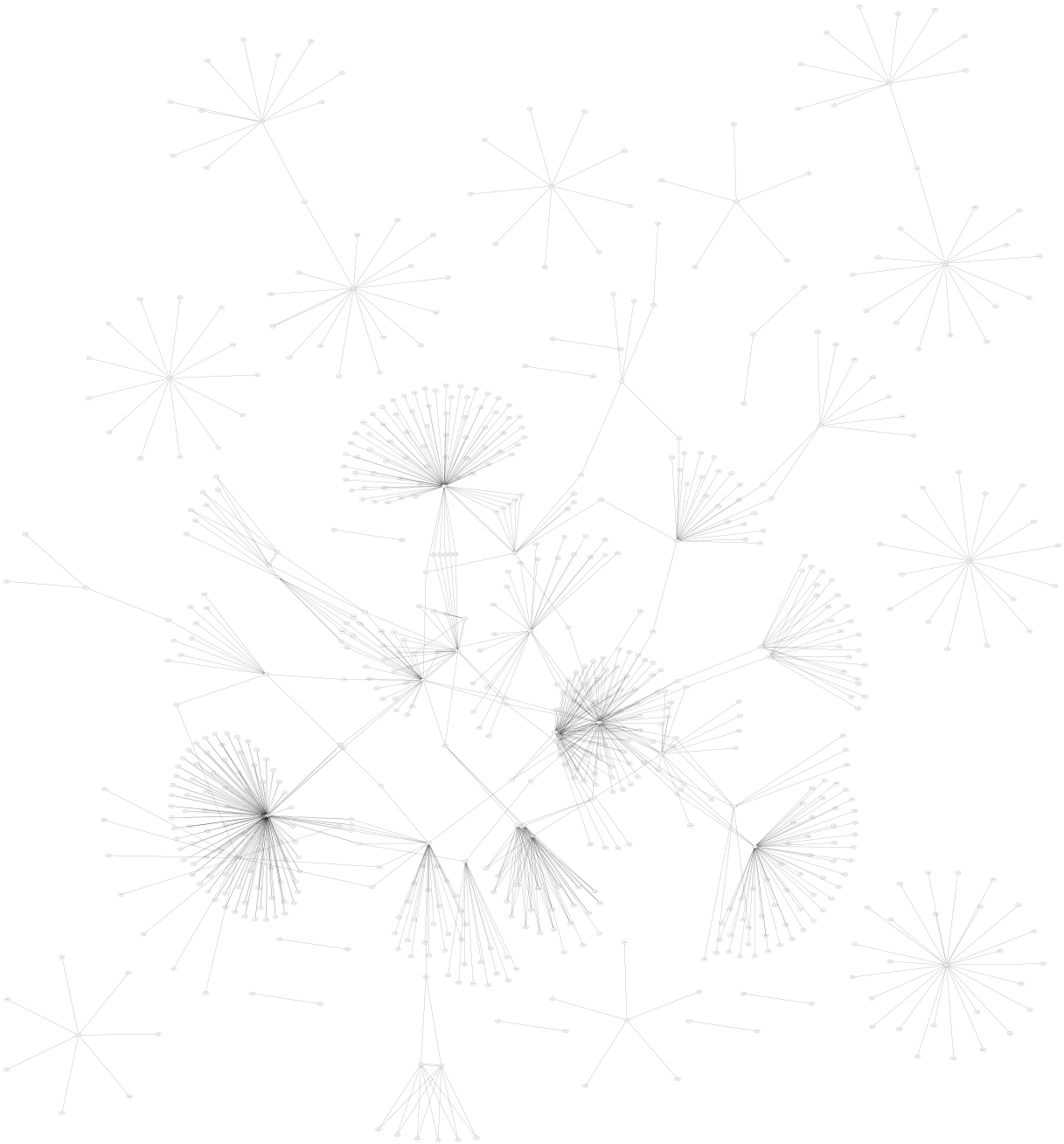

Figure 4: A subgraph of the Congressional Bills netwrok. In this network, we only illustrate 1000 edges. Hence, some documents form distinct cliques.

Table 1: Statistical results of the Congressional Bills data set.

	Number of Nodes	Number of Edges	Number of Labels	Vocabulary Size
Congressional Bills	1588	38291	1106	21007

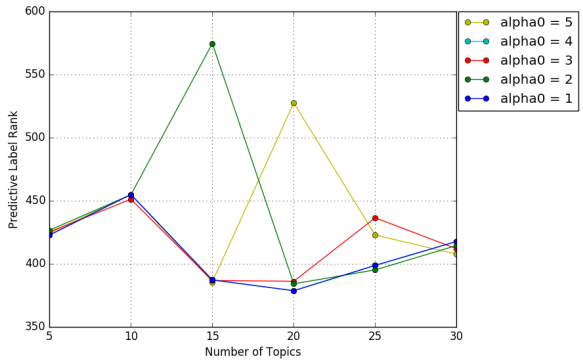

Figure 5: We change the values of $alpha0$ from 1 to 5 at intervals of 1, and the number of topics K from 5 to 30 at intervals of 5.

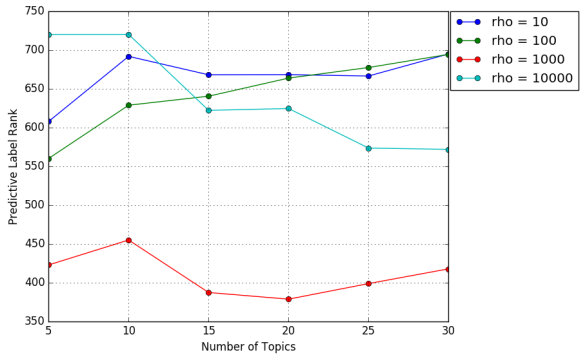

Figure 6: To tune ρ, we also change the value of it from 10, 100, 1000, to 10000.

where the link is not observed. To tune ρ, we also change the value of it from 10, 100, 1000, to 10000. As we see in the figure, again, we observe that when $K = 20$ and $\rho = 1000$, the predictive label rank is still the lowest, which means this is the best performance.

5.4 Analysis

From the results of two important parameters α and ρ, we show that when the number of topics is 20, we get the best performance. This result is consistent with the results shown in Poursabzi-Sangdeh et al. (2016). In that paper, they report that when the number of topics is 19, the topic coherence is the highest.

Topic coherence is a common way to measure the qualities of topics (Chang et al., 2009). Topic coherence is based on co-occurrence statistics, such as Normalized Pointwise Mutual Information (NPMI) (Schneider, 2005), proposed by Lau et al. (2014). Similar metrics—such as asymmetrical word pair metrics (Mimno et al., 2011) and combinations of existing measurements (Wallach et al., 2009)—correlate well with human judgments.

In their experiment, Pointwise Mutual Information (PMI) is used to evaluate the degree of coherence of topics:

$$\mathrm{PMI}(w_i) = \sum_{j}^{N-1} \log \frac{\Pr(w_i, w_j)}{\Pr(w_i)\Pr(w_j)}, \qquad (13)$$

where w_i is a topic word – the words with highest probability in that topic-word distribution, and $\Pr(w_i, w_j)$ is the co-occurrence probability of a pair of topic words based on an external corpus.

As suggested in Lau et al. (2014), when coherence score is high, it is easier for human to interpret the topic. Therefore, in the experiments in Poursabzi-Sangdeh et al. (2016), the number of topics is 19 means that the topics are easier for human to label.

In our project, we use the annotation from that setting, and use our relational topic model to test if htat is true. If topic coherence is related to human interpretability, when we use the annotated dataset to construct a network where the annotations are based on human interpretation, we should get similar results on the topic coherence. We show an illustration on how this works in Figure 7.

According to our experiment results, we confirm that when $K = 20$ the performance is the best, based on predictive label rank measurement. This result is very close to the original setting in Poursabzi-Sangdeh et al. (2016). Therefore, we conclude that their result is able to be recovered and so reliable.

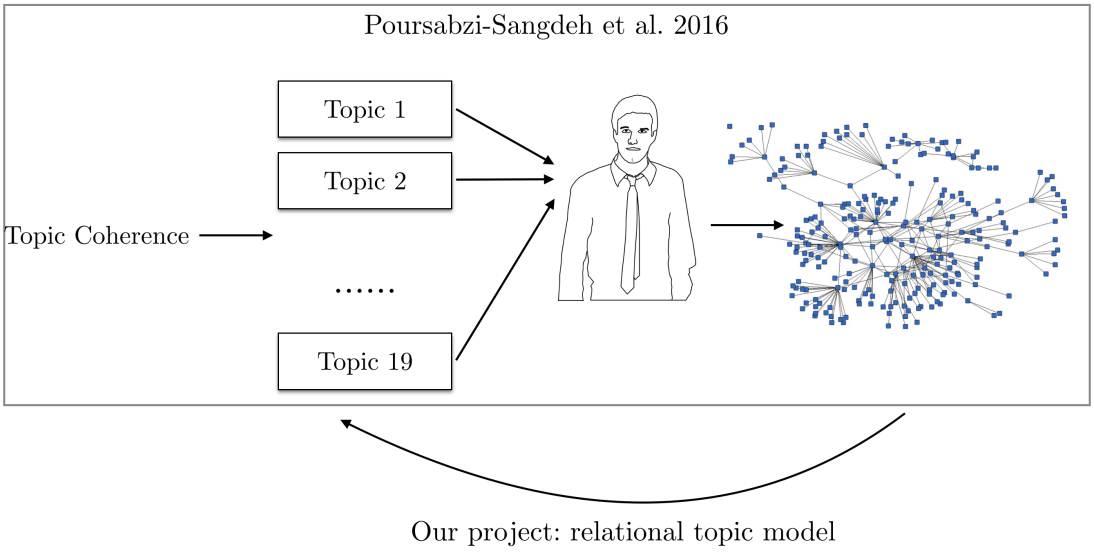

Figure 7: An illustration on how we validate the published results in our project. In Poursabzi-Sangdeh et al. (2016), they first use topic coherence, or PMI, to decide the best number of topics. Then, they speed up document annotation based on the topics, and the users completed the document labeling based on the results from topic model. In our project, however, we use the constructed network to recover the topic settings to validate if the choice of number of topics ($K = 19$ in this case) is a reasonable one.

6 Discussions

In this section, we look back to our project, and discuss some possible drawbacks in our work, and propose possible ways to improve our work.

6.1 The Construction of Network

We argue that the construction of our document network has its own drawback. First, recall that each document has its own label set from multiple users (40 in our experiments). It is very possible that it has similar labels but different names, such as *farming* v.s. *agriculture*. Thus, the label set could be very sparse. Notice that in the annotation sessions, the users could choose an existing label, so to some extend this problem could have less impact.

Second, it is also possible that some people think document d and document d' are supposed to have same labels, but some people don't. In this case, we link d and d' as long as they contain the same label. It could be improved in some wat. For example, we assign a threshold δ, so that if less than δ users assign d and d' with the same label, we ignore this link and do not add an edge between these two documents.

6.2 The Evaluation Method

Even though we use predictive label rank as our metric in this project. We found that this metric is not very accurate and reasonable at some times. For exmple, for a document d, it has three edges, i.e., (d, d_1), (d, d_2), and (d, d_3). We use two different models, i.e., M_1 and M_2, to predict links for d. The result of M_1 is $r_{d_1} = 1$, $r_{d_2} = 2$, and $r_{d_3} = 100$, which means that M_1 predicts two edges of d correctly. For model M_2, the prediction is $r_{d_1} = 1$, $r_{d_2} = 50$, and $r_{d_3} = 52$, which means that M_2 can only predict one edge of d correctly. However, when we use the predictive label rank to evaluate the performances of these two models, based on Equation 12, they have the same predictive label rank value, which indicates the same performance.

How to improve predictive label rank is an open problem now. There are manhy possibly ways to improve this measurements. For example, we feel that it may be better if we use weighted average rather than average in Equation 12. The weights could be set according to its relative positions in the ranking. Another solution could be that we put a penalty on the ranks of a document. In this way, if the predictions are $r_{d_1} = 1$, $r_{d_2} = 50$, and $r_{d_3} = 52$,

we might directly ignore r_{d_1}, and only average over r_{d_2} and r_{d_3}. Finally, we might use crowdsourcing to actively learn the quality of the network.

6.3 Implementation

As mentioned in the previous sections, we have two choice to infer the document-topic distributions θ and topic-word distributions ϕ, that is, variational inference which is used in this project, and Gibbs sampling.

The main advantages of variational inference over Gibbs sampling are that it is faster than Gibbs sampling to converge and it could be easily scaled to process huge datasets. For huge datasets which is very common in network research, like document network, or social network, variational inference is a much better method. Usually, to scale up variational inference, one needs to modify the algorithm to make it adaptive to huge datasets. In our implementation, however, we use the basic version of variational inference, since the network is not too huge to deal with using the basic version.

To scale relational topic model to very large data set, we need to use a novel method, i.e., stochastic variational inference (Hoffman et al., 2013), to train the model. Stochastic variational inference train the global parameters of the model using stochastic gradient method. At each training iteration, it only needs to analyze a subset of the corpus to form the stochastic gradient. Hence, the training speed will not be impacted by the size of the network.

7 Related works

In this section, we are going to make a brief review on the researches most related to our project.

Many efforts have been done for combining topic model and network model to analyze document networks. BKN (Zhu et al., 2013) combines the classic ideas in topic modeling with a variant of the mixed-membership block model (Airoldi et al., 2008) recently developed in the statistical physics community. This model can be inferred with a simple and scalable expectation-maximization algorithm. Hence, it can easily analyze a data set with 1.3 million words and 44 thousand links in a few minutes.

LBH-RTM (Yang et al.,) embeds a weighted stochastic block model (Aicher et al., 2014) to the relational topic model. This model can makes fuller use of the rich link structure within a document network and identify blocks in which documents are densely connected.

Markov random topic fields (Daumé III, 2009) incorporates Markov random field with LDA. This model assume that the documents which link together have the similar topic structure. It uses many factors such as shared authors and citations as edges of the network.

8 Conclusions

In this project, we implement a popular latent variable model, namely relational topic model. We also construct a new document network using Congressional Bills corpus and the user label sets generated by ALTO. We argue that our new netwok is more suitable to relational topic model, which is developed for undirected networks. In order to evaluate the correctness of the user label sets and the performance of relational topic model, we use relational topic model to fit our Congressional Bills network. Our empirical results show that when the topic number is around 20, the model is the best. This result is consistent with the previously published result.

When we look back to our project, we find some possible drawbacks. First, the method we used to construct the network may not be perfectly good. Second, predictive label rank is not very accurate at some times. Third, we implement relational topic model with variational inference, which is inefficient when the network is very large.

Our future works include first, to figure out a better way to construct the Congressional Bills network. Second, to develop a new metric for evaluating the models. Third, to implement relational topic model with stochastic variational inference. Fourth, to use the new implementation to analyze very large document networks.

References

E Scott Adler and John Wilkerson. 2006. Congressional bills project. *NSF*, 880066:00880061.

Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. 2014. Learning latent block structure in weighted networks. *Journal of Complex Networks*, page cnu026.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

Matthew J. Beal. 2003. *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University College London, UK.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003a. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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

Jonathan Chang and David M. Blei. 2009. Relational topic models for document networks. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, pages 81–88.

Jonathan Chang and David M Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, pages 124–150.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pages 288–296.

Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. 2015. Discriminative relational topic models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(5):973–986.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom M. Mitchell, Kamal Nigam, and Seán Slattery. 1998. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26-30, 1998, Madison, Wisconsin, USA.*, pages 509–516.

W. Bruce Croft, R. Wolf, and Roger Thompson. 1983. A network organization used for document retrieval. In *Research and Development in Information Retrieval, Sixth Annual International ACM SIGIR Conference, National Library of Medicine, Bethesda, Maryland, USA, June 6-8, 1983*, pages 178–188.

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

Huan Gui, Yizhou Sun, Jiawei Han, and George Brova. 2014. Modeling topic diffusion in multi-relational bibliographic information networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, pages 649–658.

Weiyu Guo, Shu Wu, Liang Wang, and Tieniu Tan. 2015. Social-relational topic model for social networks. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1731–1734.

Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Jake M Hofman and Chris H Wiggins. 2008. Bayesian approach to network modularity. *Physical review letters*, 100(25):258701.

Tomas Hrycej. 1990. Gibbs sampling in bayesian networks. *Artif. Intell.*, 46(3):351–363.

Charles Kemp, Thomas L Griffiths, and Joshua B Tenenbaum. 2004. Discovering latent classes in relational data.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 530–539.

Todd A. Letsche and Michael W. Berry. 1997. Large-scale information retrieval with latent semantic indexing. *Inf. Sci.*, 100(1-4):105–137.

Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Inf. Retr.*, 3(2):127–163.

David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 262–272.

Forough Poursabzi-Sangdeh, Jordan L. Boyd-Graber, Leah Findlater, and Kevin D. Seppi. 2016. ALTO: active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Karl-Michael Schneider. 2005. Weighted average pointwise mutual information for feature selection in text categorization. In *Knowledge Discovery in Databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005, Proceedings*, pages 252–263.

Ted H. Szymanski. 1987. Interconnection network modelling using monte carlo methods, markov chains and performance petri nets. In *Computer Performance and Reliability, Proceedings of the Second International MCPR Workshop held in Rome, Italy, May 25-29, 1987*, pages 259–274.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2012. Hierarchical dirichlet processes. *Journal of the american statistical association*.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David M. Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 1105–1112.

Weiwei Yang, Jordan Boyd-Graber, Jordan Boyd Graber, and Philip Resnik. A discriminative topic model using document network structure.

Yaojia Zhu, Xiaoran Yan, Lise Getoor, and Cristopher Moore. 2013. Scalable text and link analysis with mixed-topic link models. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 473–481. ACM.