Incorporating User Input with Topic Modeling

Yi Yang
Northwestern University
yiyang@u.northwestern.edu

Shimei Pan
University of Maryland,
Baltimore County
shimeij6@gmail.com

Jie Lu
IBM T.J.Watson Research Center
Yorktown Heights, NY
jielu@us.ibm.com

Mercan Topkara
IBM T.J.Watson Research Center
Yorktown Heights, NY
mtopkara@us.ibm.com

Doug Downey
Northwestern University
ddowney@eecs.northwestern.edu

ABSTRACT

Topic models such as Latent Dirichlet Allocation (LDA) can discover topics from a large collection of documents in an unsupervised fashion and thus is one of the most popular text analysis tool currently in use. However, when using it in practice, the topics discovered by topic model don’t always make sense to end users. The poor quality topics will substantially undermine a topic model system’s usability. Due to the unsupervised nature of topic model, this is difficult to incorporate user’s domain knowledge or feedback to the topic model. In this paper, we introduce a novel constrained LDA model, named cLDA, that is capable of incorporating user inputs in the form of document pairwise constraints. Document pairwise constraints can be document must-links and document cannot-links which represent the semantic similarity of documents. The effectiveness of the proposed cLDA model is shown in several aspects on a benchmark dataset.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms
Algorithms, Experimentation

Keywords
Topic Models, Constrained Models, User Feedback

1. INTRODUCTION

Probabilistic topic models, such as Probabilistic Latent Semantic Indexing [9], Latent Dirichlet Allocation (LDA) [5] provide a powerful framework for exploring hidden thematic patterns in text and thus are one of the most popular text analysis tool currently in use. Topic models also have had considerable applications beyond natural language processing in computer vision [7], biology [16], and psychology.

The strength of topic models is that they are unsupervised. They do not require any a priori annotations. The only input a topic model requires is the text divided into documents and the number of topics you want it to discover, and there are also models and heuristics for selecting the number of topics automatically [19].

However, when using topic models in practice, users often face one critical problem: topics discovered by the model don’t always make sense. A topic may contains thematic unrelated or incoherent words, and two thematic related words appears in two different topics. The poor quality topics frustrate users and also undermines the topic model system’s usability. Part of the problem is that topic model’s objective function doesn’t always match with human judgements [6]. Another crucial reason is that topic models, with its unsupervised and bag-of-words nature, lack of necessary information and guidance from users.

Potentially, we can solve these problems by incorporating additional user guidance or domain knowledge in topic modeling. With standard LDA however, it is impossible for users to interact with the model and provide feedback. [10] proposed an interactive topic modeling framework that allows users to add word must-links. However, it cannot handle polysemes. For example, the word “pound” can refer to either a currency or a unit of mass. If a user adds a must-link between “pound” and another financial term, then he/she cannot add a must-link between “pound” and any measurement terms. Since word must-links are added without context, there is no way to disambiguate them. As a result, word constraints frequently are not as effective as document constraints.

In this paper, we investigate another type of user inputs, document level feedbacks. So far, there isn’t any existing LDA method that can handle both document must-links and cannot-links constraints effectively. We developed a novel Constrained LDA algorithm (cLDA) which allows LDA to incorporate pairwise document constraints in the form of document must-links and cannot-links. Our evaluation on a benchmark dataset has demonstrated the effectiveness of
cLDA in incorporating these constraints.

2. RELATED WORK

Our work is closely related to the field of topic modeling with prior knowledge. Recently, new algorithms have been proposed to extend the standard LDA algorithm to incorporate additional user feedback or domain knowledge. Among them, some support seed constraints (e.g., document topic labels) [4, 14] while others support pairwise word constraints (e.g., word must-links and cannot-links) [1, 2, 10]. To ensure topic stability and at the same time to avoid over-constraining the new topic model, it is desirable to encode the topic stability constraints as pairwise document constraints (e.g., document must-link and cannot-link constraints). Currently, without further extension, none of these methods supports general pair-wise document constraints.

There is also work focusing on incorporating specific link relations (e.g., citations between documents) to facilitate the learning of the topic model. In such a way, we use an asymmetric Dirichlet prior over topic-word symmetric [20]. In the following, we explain how to incorporate the prior knowledge encoded in document must-links and cannot-links into an LDA topic model.

3. CONSTRAINED TOPIC MODEL

We denote the collection of documents as \( \mathcal{D} = \{d_1, d_2, ..., d_N\} \). As in LDA [5] we denote \( \bar{\theta}_d \) as the topic distribution of a document \( d \). Let \( \bar{\alpha}_g = (\alpha_{g,1}, ..., \alpha_{g,K}) \) be the hyper-parameter of the Dirichlet prior over \( \theta_d \), where \( K \) is the number of topics.

To incorporate prior knowledge, we focus on two types of document level constraints: must-links and cannot-links. A must-link constraint indicates that two documents should share the same topics (e.g., between two Sports articles), and a cannot-link constraint indicates that two documents should have different topics (e.g., between a Sports and a Politics article). For example, Table 1 shows the snippets of four documents. We can see that \( \text{doc1} \), \( \text{doc2} \) and \( \text{doc4} \) are about Sports articles that the first one is about hockey and the rest two are about baseball. Also, \( \text{doc3} \) is a space topic article. Therefore, from our definition on must-link and cannot-link, users can add must-link between \{\text{doc2}, \text{doc4}\} and add cannot-link between \{\text{doc1}, \text{doc3}\}, \{\text{doc2}, \text{doc3}\} and \{\text{doc3}, \text{doc4}\}. However, different users may have different opinion on hockey topic and baseball topic articles. Some may say both are about Sport topic so there should be a must-link between them, while others who look for more fine-grained topics may argue they are two different topics so they share cannot-link relation. Both arguments reflect users understanding of the documents. Therefore, we hope that constrained LDA can take constraints as soft preferences rather than hard constraints.

We denote \( \mathcal{M}_i \), \( \mathcal{C}_i \) as the set of documents sharing must-links with document \( d_i \), and \( \mathcal{C}_i \), \( \mathcal{C}_i \) as the set of documents sharing cannot-links with document \( d_i \).

3.1 The Role of Concentration Parameters

In LDA, the prior of \( \bar{\theta} \) is a Dirichlet distribution, which is denoted by \( \text{Dir} (\bar{\alpha}_g) = \text{Dir} (\alpha_{g,1}, ..., \alpha_{g,K}) \). \( \bar{\alpha}_g \) is the global hyperparameter, and \( \alpha_{g,i} \) determines how “concentrated” the probability mass of a sampled \( \bar{\theta} \) is likely to be on topic \( i \). If all the \( \alpha_{g,i} \) are less than one, the probability mass tends to be concentrated on a few topics. If all the \( \alpha_{g,i} \) are greater than one, the probability mass tends to be more uniformly distributed. Meanwhile, smaller \( \alpha_{g,i} \) attracts more concentration on topic \( i \). A simple and commonly used Dirichlet distribution is the symmetric Dirichlet distribution, where \( \alpha_{g,1} = \alpha_{g,2} = \ldots = \alpha_{g,K} = \alpha \).

If there is no prior knowledge, the concentration parameters are normally set to be equal. Then the model can update the likelihood by learning the posterior of \( \bar{\theta} \). When we have knowledge about a document’s topic distribution, e.g., two documents have similar topic distribution, the topic distributions of documents cannot be assumed to be independently sampled. To achieve this, we manipulate the Dirichlet prior over document-topic distribution such that document must-link and cannot-link constraints can be incorporated into the topic model. In such a way, we use an asymmetric Dirichlet prior over document-topic while keep the Dirichlet prior over topic-word symmetric [20]. In the following, we explain how to incorporate the prior knowledge encoded in document must-links and cannot-links into an LDA topic model.

3.2 Must-link Constraint

A must-link between two documents \( d_1 \) and \( d_2 \) suggests that \( d_1 \) and \( d_2 \) should share the same topics, e.g., both are Sports news. Thus \( \bar{\theta}_1 \) should be similar to \( \bar{\theta}_2 \). If we project both distributions to a simplex, they should be close to each other. See \( x_1 \) ad \( x_2 \) in Figure 1(b).

Given the documents in \( \mathcal{M}_i \), we introduce an auxiliary variable \( \bar{\alpha}_{i}^{M} \):

\[
\bar{\alpha}_{i}^{M} = T \ast \frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} \bar{\theta}_j, \tag{1}
\]

where \( T \) controls the concentration parameters. The larger the value of \( T \) is, the closer \( \bar{\theta}_i \) is to the average of \( \bar{\theta}_j \)’s. \( \mathcal{M}_i \in \mathcal{D} \) is the set of documents sharing must-links with document \( d_i \).

3.3 Cannot-link Constraint

A cannot-link between documents \( d_1 \) and \( d_2 \) suggests that \( d_1 \) and \( d_2 \) should not have the same topics, for example, one is about Sports and the other is about Politics. Thus, \( \bar{\theta}_1 \) should not be similar to \( \bar{\theta}_2 \). If we project both distributions to a simplex, they should be really far away from each other. See \( x_1 \) ad \( x_3 \) in Figure 1(b).

Given the documents in \( \mathcal{C}_i \), we introduce the following auxiliary variable:

\[
\bar{\alpha}_{i}^{C} = T \ast \arg\max_{j \in \mathcal{C}_i} \min_{j \notin \mathcal{C}_i} KL(\bar{\theta}_i, \bar{\theta}_j), \tag{2}
\]

where \( KL(\bar{\theta}_i, \bar{\theta}_j) \) is the KL-divergence between two distributions \( \bar{\theta}_i \) and \( \bar{\theta}_j \). This means we choose a vector that is maximally far away from \( \mathcal{C}_i \), in terms of KL divergence to
in Salt Lake City this past Sunday, the local ABC station decided not to televise the **hockey** games...

...Hello, my friends and I are running the Homewood Fantasy Baseball League (pure fantasy **baseball** teams)...

...According the IAU Circular #5744, Comet Shoemaker-Levy 1993e, may be temporarily in orbit around **Jupiter**...

...He has obtained the play by play records for every major league **baseball** game for the past several years...

Table 1: Four document snippets selected from 20 Newsgroup dataset.

<table>
<thead>
<tr>
<th>Document</th>
<th>Snippet</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc1</td>
<td>...in Salt Lake City this past Sunday, the local ABC station decided</td>
</tr>
<tr>
<td></td>
<td>not to televise the <strong>hockey</strong> games...</td>
</tr>
<tr>
<td>doc2</td>
<td>...Hello, my friends and I are running the Homewood Fantasy Baseball</td>
</tr>
<tr>
<td></td>
<td>League (pure fantasy <strong>baseball</strong> teams)</td>
</tr>
<tr>
<td>doc3</td>
<td>...According the IAU Circular #5744, Comet Shoemaker-Levy 1993e, may</td>
</tr>
<tr>
<td></td>
<td>be temporarily in orbit around <strong>Jupiter</strong>...</td>
</tr>
<tr>
<td>doc4</td>
<td>...He has obtained the play by play records for every major league</td>
</tr>
<tr>
<td></td>
<td><strong>baseball</strong> game for the past several years...</td>
</tr>
</tbody>
</table>

its nearest neighbor in $C_i$. $C_i \in \mathcal{D}$ is the set of documents sharing cannot-links with document $d_i$.

Then in each iteration, we draw a $\tilde{\theta}_i$ from the following distribution:

$$
\tilde{\theta}_i \sim \text{Dir}(\eta_g \alpha_g + \eta_M \alpha_i^M + \eta_C \alpha_i^C) = \text{Dir}(\tilde{\alpha}_i).
$$

Here, $\eta_g$, $\eta_M$ and $\eta_C$ are the weights to control the trade-off among the three terms. Note that in the first iteration of learning, it is possible that all the $\theta_i$'s are initialized by drawing from the global $\alpha_i$ solely. In our experiment, we choose $T = 100$, $\eta_g = \eta_M = \eta_C = 1$.

Figure 1(a) shows the graphical model of cLDA. In this example, we assume $d_1$ and $d_2$ are in must-link relation, and $d_2$ and $d_3$ are in cannot-link relation. $d_4$ is not in any relation with the other documents. When projecting their topic distributions to a three-topic simplex, the distance between two documents reveals their relations.

![Figure 1](http://example.com/figure1.png)

Figure 1: (a) The graphical model of cLDA. (b) The $\theta$s of four documents plotted on a simplex.

3.4 Inference with Gibbs Sampling

Given a set of document must-links $\mathcal{M}$ and cannot-links $\mathcal{C}$, we infer the values of the hidden variables $z$ using collapsed Gibbs sampling as in LDA [8]. In each iteration, a topic assignment $z_{i,j}$ of word $w_{i,j}$ in document $d_i$ is sampled based on all the other variables according to the following distribution:

$$
p(z_{i,j} = t | w_{i,j}, z_{-i,j}, \tilde{\alpha}_g, \tilde{\beta}, M_i, C_i) = 
\frac{n_{t,k}^{+} + \eta_t}{\sum_{t=1}^{V} n_{t,k}^{+} + \eta_t^{+}}
$$

We simulate user inputs using the documents' ground truth labels. If two documents have the same label, we add a must-link between them. Similarly, we also add a cannot-link between two documents with different labels. All these constraints are added into a constraint pool.

Table 2: Statistics of three sub-datasets.

<table>
<thead>
<tr>
<th>Sub-dataset</th>
<th>Sim3</th>
<th>Diff3</th>
<th>Mix3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{train}$</td>
<td>1,768</td>
<td>1,670</td>
<td>1,790</td>
</tr>
<tr>
<td>$D_{test}$</td>
<td>1,178</td>
<td>1,110</td>
<td>1,190</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>4,222</td>
<td>4,822</td>
<td>4,685</td>
</tr>
</tbody>
</table>

1Available at [http://people.csail.mit.edu/jrennie/20Newsgroups](http://people.csail.mit.edu/jrennie/20Newsgroups)
To verify that with additional document constraints cLDA in blue and space in yellow.

cannot-links. Here, hockey documents are in red, baseball with cannot-links; d)cLDA with both must-links and LDA with no constraints; b)cLDA with must-links; c)cLDA links are clustered together while the θ

deferred by standard LDA and cLDA. For the keywords derived by both LDA and
cLDA, we performed a post-processing step to re-rank them using [17]. As shown in Table 3, standard LDA had more
trouble learning the keywords related to the Baseball topic. cLDA, however, successfully separated all three topics, and
the keywords representing each topic are quite coherent and informative.

4.2 Experiment 2: Topic Coherence
In the second experiment, we evaluate cLDA on topic coherence. Recent research [12] has shown that topic coherence is highly consistent with human judgement than perplexity. Thus, here we use topic coherence to assess a topic model’s quality. Following [12], topic t’s coherence is defined as

\[ C(t : V(t)) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v^{(t)}, v^{(t)})}{D'(v^{(t)}, v')} \]

where \( D(v) \) is the document frequency of word type \( v \), \( D(v, v') \) is the co-document frequency of word type \( v \) and \( v' \), and \( V(t) = (v_1^{(t)}, ..., v_M^{(t)}) \) is a list of the \( M \) most probable words in topic \( t \). In our experiments, we choose the 20 most probable words to compute topic coherence, i.e., \( M = 20 \). In addition, since LDA usually generates common background topics which appear in many documents and thus un-interesting, we also filtered those topics based on the method proposed in [17] before we compute the coherence scores for all the methods. A higher topic coherence indicates a higher quality of topics.

\[ \phi_{t} = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v^{(t)}, v^{(t)})}{D'(v^{(t)}, v')} \]

As shown in Figure 3, cLDA with both must-link and cannot-link constraints achieves a better topic coherence score than the baseline LDA on Mix3 dataset. Particularly, it improves the topic coherence score by 14.9% than LDA. Note that we ran the experiments 10 times, and the topic coherence score are the average number and the results are statistically significant at \( p = 0.05 \).

4.3 Experiment 3: Document Classification
In this experiment, we evaluate the effectiveness of cLDA on document classification. For this experiment, we developed a multi-class classifier. First, given the training documents in each newsgroup, we computed the centroid of their \( \theta_s \). This centroid captures the general topic mixture of a particular newsgroup. During classification, for each document in the testing set, we compute the KL divergence of the \( \theta \) of
Table 3: Top 10 keywords of each topic by LDA (above) and cLDA (below).

<table>
<thead>
<tr>
<th>Topic</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseball</td>
<td>writes think good know better even baseball really going anyone</td>
</tr>
<tr>
<td>space</td>
<td>space nasa launch orbit satellite moon earth data mission shuttle</td>
</tr>
<tr>
<td>hockey</td>
<td>game hockey playoff division pittsburgh run detroit canada boston pick</td>
</tr>
<tr>
<td>baseball</td>
<td>baseball run base thanks brave pitching cub ball hitter yankee</td>
</tr>
<tr>
<td>space</td>
<td>space nasa system launch orbit satellite moon science center earth</td>
</tr>
<tr>
<td>hockey</td>
<td>play hockey goal playoff period pittsburgh leaf wing detroit ranger</td>
</tr>
</tbody>
</table>

The new document and each of the topic centroid. The topic label of the testing document is determined by the news-group with the smallest KL distance. Finally, we compare the derived labels with the ground-truth. Table 4 shows how cLDA can help reduce classification error. On all three datasets, cLDA with must-links or cannot-links or both can significantly reduce classification errors. In addition, the total relatively classification error reduction by incorporating both must-links and cannot-links for Diff3, Sim3 and Mix3 are 14.3%, 27.5% and 51.9% respectively.

<table>
<thead>
<tr>
<th>Table 4: Classification error rate.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Model</td>
</tr>
<tr>
<td>LDA</td>
</tr>
<tr>
<td>cLDA with only must-links</td>
</tr>
<tr>
<td>cLDA with only cannot-links</td>
</tr>
<tr>
<td>cLDA with both links</td>
</tr>
<tr>
<td>Error Reduction by cLDA with both links</td>
</tr>
</tbody>
</table>

4.4 Experiment 4: Constraint Sampling

For a corpus which has a large number of document annotations, the number of constraints would be huge if we exhaustively generate all the pairwise must-link and cannot-link constraints. Fortunately, in each iteration, we only need to randomly sample a small subset of the constraints. Figure 4 shows that the model converged quickly when only a limited number of constraints were sampled in each iteration. For example, on both Mix3 and Sim3, the model quickly converged to a stationary error rate around 9% and 38% respectively after only 30 constraints were sampled per iteration. For Diff3, only 50 constraints were sampled in each iteration before the model converged to a stationary error rate around 8%. Please note that for all the models in Experiment 2 and 3, we report the average over 10 random runs.

5. DISCUSSION

5.1 Compare with Seed Constraint

In the above sections, we discuss the method of incorporating document pairwise constraints into topic model. We can therefore acquire user feedback or domain expert’s prior knowledge in the form of pairwise document constraints. There is another type of constraints that are often used in practice, which is document seed constraint. A document seed constraint restricts a document to be about a set of predefined topics, as in Labeled LDA [14]. However, seed constraints are not easy to obtain in topic model settings since it is impractical to list existing topics before training the model. On the other hand, seed constraint has one natural limitation, which is that users might assign different word tags to represent the same meaning. For example, for a Fifa wordcup related article, one user might give it “soccer” tag while the other user might use “football” instead. This multiple-to-one word tag to topic mapping will break the assumption of Labeled LDA, which assumes a one-to-one word tag to topic mapping. Therefore, we believe that pairwise constraints are more practical than seed constraint for user interaction in practical. We will conduct more comprehensive experiments to demonstrate the superior of pairwise constraint to seed constraint in the future.

5.2 Acquire User Input

User inputs can be acquired before or during the model’s training. If we obtain user inputs from domain experts in advance, then these prior knowledge will be used in the training. However, this might not be the best way to acquire user inputs since some of the user inputs might not be informative to the model at all. For example, if the model can successfully distinguish two documents very well, we don’t have to ask user to label these two documents before the
training. Active learning [15] provides a useful framework which allows users to iteratively give feedback to the model to improve its quality. In general, with the same amount of human labeling, active learning often results in a better model than that learned by an off-line method. Interactively acquiring User inputs and incorporating the inputs into topic modeling will be a future research direction.

6. CONCLUSION

Topic model such as LDA is one of the most popular text analysis tools currently in use. However, due to its unsupervised nature, users cannot easily incorporate their domain knowledge or feedback into the model. In this paper, we introduce an LDA extension cLDA, a model that can directly encode user inputs with LDA, in the form of document pairwise constraints. Our preliminary results show that cLDA outperforms LDA in classification error reduction and topic coherence metrics. Document pairwise constraints are more easily to obtain than document seed constraints in the topic modeling circumstances, and we also show the potential usage of interactive topic modeling with the framework of cLDA.

7. REFERENCES