Active Learning with Constrained Topic Model

Yi Yang
Northwestern University
yiyang@u.northwestern.edu

Shimei Pan
IBM T. J. Watson Research Center
shimei@us.ibm.com

Doug Downey
Northwestern University
ddowney@eecs.northwestern.edu

Kunpeng Zhang
University of Illinois at Chicago
kzhang6@uic.edu

Abstract

Latent Dirichlet Allocation (LDA) is a topic modeling tool that automatically discovers topics from a large collection of documents. It is one of the most popular text analysis tools currently in use. In practice however, the topics discovered by LDA do not always make sense to end users. In this extended abstract, we propose an active learning framework that interactively and iteratively acquires user feedback to improve the quality of learned topics. We conduct experiments to demonstrate its effectiveness with simulated user input on a benchmark dataset.

1 Introduction

Statistical topic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) provide powerful tools for uncovering hidden thematic patterns in text and are useful for representing and summarizing the contents of large document collections. However, when using topic models in practice, users often face one critical problem: topics discovered by the model do not always make sense. A topic may contain thematically unrelated words. Moreover, two thematic related words may appear in different topics. This is mainly because the objective function optimized by LDA may not reflect human judgments of topic quality (Boyd-Graber et al., 2009).

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. (Hu et al., 2011) proposed an interactive topic modeling framework that allows users to add word must-links. However, it has several limitations. Since the vocabulary size of a large document collection can be very large, users may need to annotate a large number of word constraints for this method to be effective. Thus, this process can be very tedious. More importantly, 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.

Active learning (Settles, 2010) 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.

In this extended abstract, we propose an active learning framework for LDA. It is based on a new constrained topic modeling framework which is capable of handling pairwise document constraints. We present several design choices and the pros and cons of each choice. We also conduct simulated experiments to demonstrate the effectiveness of the approach.

2 Active Learning With Constrained Topic Modeling

In this section, we first summarize our work on constrained topic modeling. Then, we introduce an active topic learning framework that employs constrained topic modeling.

In LDA, a document's topic distribution $\vec{\theta}$ is drawn from a Dirichlet distribution with prior $\vec{\alpha}$. A simple and commonly used Dirichlet distribution uses a symmetric $\vec{\alpha}$ prior. However, (Wallach et al., 2009) has shown that an asymmetric Dirichlet prior over the document-topic distributions $\vec{\theta}$ and a symmetric Dirichlet prior over the topic-word distributions $\vec{\phi}$ yield significant improvements in model performance. Our constrained topic model uses asymmetric priors to encode constraints.

To incorporate user feedback, we focus on two

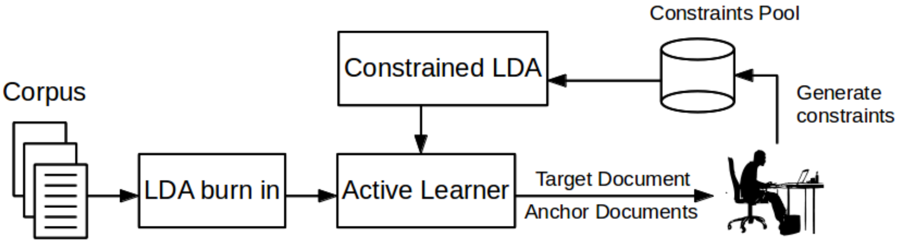

Figure 1: Diagram illustrating the topic model active learning framework.

types of document constraints. A **must-link** between two documents indicates that they belong to the same topics, while a **cannot-link** indicates that they belong to different topics.

Previously, we proposed a constrained LDA framework called cLDA,[1] which is capable of incorporating pairwise document constraints. Given pairwise document constraints, the topic distribution of a document cannot be assumed to be independently sampled. More specifically, we denote the collection of documents as $\mathcal{D} = \{d_1, d_2, ..., d_N\}$. We also denote $\mathcal{M}_i \in \mathcal{D}$ as the set of documents sharing must-links with document d_i, and $\mathcal{C}_i \in \mathcal{D}$ as the set of documents sharing cannot-links with document d_i. $\vec{\theta}_i$ is the topic distribution of d_i, and $\vec{\alpha}$ is the global document-topic hyper-parameter shared by all documents.

Given the documents in $\mathcal{M}_i$, we introduce an auxiliary variable $\vec{\alpha}_i^{\mathcal{M}}$:

$$\vec{\alpha}_i^{\mathcal{M}} = T * \frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} \vec{\theta}_j, \qquad (1)$$

where T controls the concentration parameters. The larger the value of T is, the closer $\vec{\theta}_i$ is to the average of $\vec{\theta}_j$'s.

Given the documents in $\mathcal{C}_i$, we introduce another auxiliary variable:

$$\vec{\alpha}_i^{\mathcal{C}} = T * \arg_{\vec{\theta}_i} \max \min_{j \in \mathcal{C}_i} KL(\vec{\theta}_i, \vec{\theta}_j), \qquad (2)$$

where $KL(\vec{\theta}_i, \vec{\theta}_j)$ is the KL-divergence between two distributions $\vec{\theta}_i$ and $\vec{\theta}_j$. This means we choose a vector that is maximally far away from $\mathcal{C}_i$, in terms of KL divergence to its nearest neighbor in $\mathcal{C}_i$.

In such a way, we force documents sharing must-links to have similar topic distributions while documents sharing cannot-links to have dissimilar topic distributions. Note that it also encodes constraint as soft preference rather than hard constraint. We use Collapsed Gibbs Sampling for LDA inference. During Gibbs Sampling, instead of always drawing $\vec{\theta}_i$ from $Dirichlet(\vec{\alpha})$, we draw $\vec{\theta}_i$ based on the following distribution:

$$\vec{\theta}_i \sim Dir(\eta\vec{\alpha} + \eta_{\mathcal{M}}\vec{\alpha}_i^{\mathcal{M}} + \eta_{\mathcal{C}}\vec{\alpha}_i^{\mathcal{C}}) = Dir(\vec{\alpha}_i). \quad (3)$$

Here, η_g, $\eta_{\mathcal{M}}$ and $\eta_{\mathcal{C}}$ are the weights to control the trade-off among the three terms. In our experiment, we choose $T = 100$, $\eta_g = \eta_{\mathcal{M}} = \eta_{\mathcal{C}} = 1$.

Our evaluation has shown that cLDA is effective in improving topic model quality. For example, it achieved a significant topic classification error reduction on the 20 Newsgroup dataset. Also, topics learned by cLDA are more coherent than those learned by standard LDA.

2.1 Active Learning with User Interaction

In this subsection, we present an active learning framework to iteratively acquire constraints from users. As shown in Figure 1, given a document collection, the framework first runs standard LDA with a burnin component. Since it uses a Gibbs sampler (Griffiths and Steyvers, 2004) to infer topic samples for each word token, it usually takes hundreds of iterations for the sampler to converge to a stable state. Based on the results of the burnt-in model, the system generates a target document and a set of anchor documents for a user to annotate. Target document is a document on which the active learner solicits user feedback, and anchor documents are representatives of a topic model's latent topics. If a large portion of the word tokens in a document belongs to topic i, we say the document is an *anchor* document for topic i.

A user judges the content of the target and the anchor documents and then informs the system whether the target document is similar to any of the anchor documents. The user interface is designed so that the user can drag the target document near an anchor document if she considers both to be the same topic. Currently, one target document can be must-linked to only one anchor document. Since it is possbile to have multiple topics in one document, in the future, we will allow user to add must links between one target and mulitple anchor documents. After adding one or more must-links, the

[1] currently in submission.

system automatically adds cannot-links between the target document and the rest anchor documents.

Given this input, the system adds them to a constraint pool. It then uses cLDA to incorporate these constraints and generates an updated topic model. Based on the new topic model, the system chooses a new target document and several new anchor documents for the user to annotate. This process continues until the user is satisfied with the resulting topic model.

How to choose the target and anchor documents are the key questions that we consider in the next subsections.

2.2 Target Document Selection

A target document is defined as a document on which the active learner solicits user feedback. We have investigated several strategies for selecting a target document.

Random: The active learner randomly selects a document from the corpus. Although this strategy is the simplest, it may not be efficient since the model may have enough information about the document already.

MaxEntropy: The entropy of a document d is computed as $H_d = -\sum_{i=1}^{K} \theta_{dk} \log \theta_{dk}$, where K is the number of topics, and θ is model's document-topic distribution. Therefore, the system will select a document about which it is most confused. A uniform θ implies that the model has no topic information about the document and thus assigns equal probability to all topics.

MinLikelihood: The likelihood of a document d is computed as $L_d = (\sum_{i=1}^{N} \sum_{k=1}^{K} \phi_{ki}\theta_{dk})/N$, where N is the number of tokens in d, and ϕ is model's topic-word distribution. Since the overall likelihood of the input documents is the objective function LDA aims to maximize, using this criteria, the system will choose a document that is most difficult for which the current model achieves the lowest objective score.

2.3 Anchor Documents Selection

Given a target document d, the active learner then generates one or more anchor documents based on the target document's topic distribution θ_d. It filters out topics with trivial value in θ_d and extracts an anchor topic set T_{anc} which only contains topics with non-trivial value in θ_d. A trivial θ_{di} means that the mass of ith component in θ_d is neglectable, which indicates that the model rarely assign topic i to document d. For each topic t in T_{anc}, the active learner selects an anchor document who has minimum Euclidean distance with an ideal anchor θ'_t. In the ideal anchor θ'_t, all the components are zero except the value of the t_{th} component is 1. For example, if a target document d's θ_d is {0.5, 0.3, 0.03, 0.02, 0.15} in a $K = 5$ topic model, the active learner would generate $T_{anc} = \{0, 1, 4\}$ and for each t in T_{anc}, an anchor document.

However, it is possible that some topics learned by LDA are only "background" topics which have significant non-trivial probabilities over many documents (Song et al., 2009). Since background topics are often uninteresting ones, we use a weighted anchor topic selection method to filter them. A weighted k_{th} component of θ'_{dk} for document d is defined as follows: $\theta'_{dk} = \theta_{dk}/\sum_{i=0}^{D} \theta_{ik}$. Therefore, instead of keeping the topics with non-trivial values, we keep those whose weighted values are non-trivial.

3 Evaluation

In this section, we evaluate our active learning framework. Topic models are often evaluated using perplexity on held-out test data. However, recent work (Boyd-Graber et al., 2009; Chuang et al., 2013) has shown that human judgment sometimes is contrary to the perplexity measure. Following (Mimno et al., 2011), we employ Topic Coherence, a metric which was shown to be highly consistent with human judgment, to measure a topic model's quality. It relies upon word co-occurrence statistics within documents, and does not depend on external resources or human labeling.

We followed (Basu et al., 2004) to create a `Mix3` sub-dataset from the 20 Newsgroups data[2], which consists of two newsgroups with similar topics (rec.sport.hockey, rec.sport.baseball) and one with a distinctive topic (sci.space). We use this dataset to evaluate the effectiveness of the proposed framework.

3.1 Simulated Experiments

We first burn-in LDA for 500 iterations. Then for each additional iteration, the active learner generates one query which consists of one target document and one or more anchor documents. We simulate user feedback using the documents' ground truth labels. If a target document has the same label as one of the anchor documents, we add a must-link between them. We also add cannot-links between the target document and the rest of the anchor documents. All these constraints are added into a constraint pool. We also augment the constraint pool with derived constraints. For example, due to transitivity, if there is a must-link between (a, b) and (b, c), then we add

[2] Available at `http://people.csail.mit.edu/jrennie/20Newsgroups`

Topic	Words
1	writes, like, think, good, know, better, even, people, run, hit
2	space, nasa, system, gov, launch, orbit, moon, earth, access, data
3	game, play, hockey, season, league, fun, wing, cup, shot, score
1	baseball, hit, won, shot, hitter, base, pitching, cub, ball, yankee
2	space, nasa, system, gov, launch, obit, moon, earth, mission, shuttle
3	hockey, nhl, playoff, star, wing, cup, king, detroit, ranger

Table 1: Ten most probable words of each topic before (above) and after active learning (below).

a must link between (a, c). We simulate the process for 100 iterations to acquire constraints. After that, we keep cLDA running for 400 more iterations with the acquired constraints until it converges.

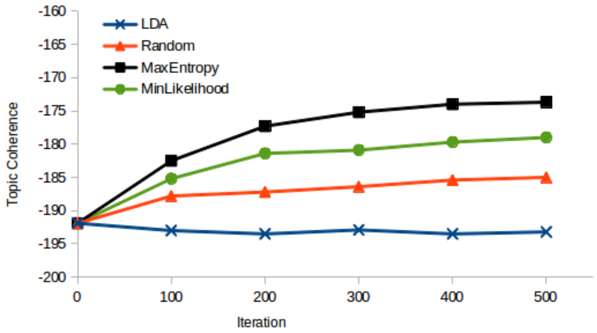

Figure 2: Topic coherence with different number of iterations.

Figure 2 shows the topic coherence scores for different target document selection strategies. This result indicates 1). MaxEntropy has the best topic coherence score. 2). All active learning strategies outperform standard LDA, and the results are statistically significant at $p = 0.05$. With standard LDA, 500 more iterations without any constraints does not improve the topic coherence. However, by active learning with cLDA for 500 iterations, the topic coherences are significantly improved.

Using MaxEntropy target document selection method, we demonstrate the improvement of the most probable topic keywords before and after active learning. Table 1 shows that before active learning, topic 1's most probable words are incoherent and thus it is difficult to determine the meaning of the topic . After active learning, in contrast, topic 1's most probable words become more consistent with a "baseball" topic. This example suggests that the active learning framework that interactively and iteratively acquires pairwise document constraints is effective in improving the topic model's quality.

4 Conclusion

We presented a novel active learning framework for LDA that employs constrained topic modeling to actively incorporate user feedback encoded as pairwise document constraints. With simulated user input, our preliminary results demonstrate the effectiveness of the framework on a benchmark dataset. In the future, we will perform a formal user study in which real users will interact with the system to iteratively refine topic models.

Acknowledgments

This work was supported in part by DARPA contract D11AP00268.

References

Sugato Basu, A. Banjeree, ER. Mooney, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *SDM*, pages 333–344.

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

Jordan Boyd-Graber, Jonathan Chang, Sean Gerrish, Chong Wang, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*.

Jason Chuang, Sonal Gupta, Christopher D. Manning, and Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *ICML*.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Yuening Hu, Jordan Boyd-Graber, and Brianna Satinoff. 2011. Interactive topic modeling. In *ACL*, pages 248–257.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *EMNLP*, pages 262–272.

Burr Settles. 2010. Active learning literature survey. Technical report, University of Wisconsin Madison.

Yangqiu Song, Shimei Pan, Shixia Liu, Michelle X. Zhou, and Weihong Qian. 2009. Topic and keyword re-ranking for lda-based topic modeling. In *CIKM*, pages 1757–1760.

Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *NIPS*, pages 1973–1981.