Supervised HDP using Prior Knowledge

Boyi Xie and Rebecca J. Passonneau

Columbia University, Center for Computational Learning Systems,
475 Riverside Drive MC 7717, 10115 New York, USA
{xie,becky}@cs.columbia.edu

Abstract. End users can find topic model results difficult to interpret and evaluate. To address user needs, we present a semi-supervised hierarchical Dirichlet process for topic modeling that incorporates user-defined prior knowledge. Applied to a large electronic dataset, the generated topics are more fine-grained, more distinct, and align better with users' assignments of topics to documents.

Keywords: topic modeling, hierarchical Dirichlet process, supervised learning

1 Introduction

Topic modeling, a method to discover semantic themes that permeate large collections of electronic documents provides a high level view of content in each document and over the collection. It is typically unsupervised, scales well, and can be done with little text pre-processing. Typically, however, only some of the topics look meaningful to end users, and it is unclear how useful a given topic model might be for improving user access to a collection. We introduce an approach to topic modeling that sacrifices some predictive power to incorporate an independent knowledge source.

This paper proposes a method to incorporate into topic modeling semantic categories of interest to users, and allows the user to control the degree to which this prior knowledge supervises the learning. Our method incorporates *a priori* semantic categories in two ways. The categories are used to initialize the set of topics, thus corresponding to topic labels. These categories are also used to label words in the vocabulary that have an *a priori* association with the categories.

We are collaborating with university librarians on a web archive of human rights sites in the use of a controlled vocabulary from the library domain, and a digital library collection. Their goals are to facilitate research on human rights websites, and to preserve sites at risk of being taken down. For subject indexing they use Library of Congress Subject Headings (LCSH), which are an integral part of bibliographic control. We demonstrate the use of a set of human rights LCSH terms in our method, and investigate how librarians and non-librarians assign the generated topics to documents in the collection.

Sections 2 (related work) and 3 (methods) provide context for the experiments described in section 4. We contrast the number and distinctness of topics in results from standard (unsupervised) hierarchical Dirichlet process (HDP) and three levels of supervised HDP. We compare unsupervised to lightly supervised models in our user study. In the user study results, topics from the supervised topic model align more closely with librarians' assignments of topics to websites.

2 Related Work

[4] first introduce latent Dirichlet allocation (LDA) to topic modeling, building on the idea from Latent Semantic Analysis (LSA) that hidden semantic dimensions condition the distribution of words in documents. LDA replaces matrix decomposition with a probabilistic model. [10] discuss the hierarchical Dirichlet process (HDP) and use a non-parametric Bayesian model.

Previous work on topic modeling has investigated the introduction of supervision. The models (e.g. [3], [8]) are based on document labels while we supervise using word characteristics. Labeled LDA in [9] that rely on supervision on words are relevant to our study, and z-label in [2] reflects a similar idea of domain dependent modeling. However, all these are LDA models while ours is based on hierarchical Dirichlet process, which is a non-parametric Bayesian approach. As such, it is more readily extended to supervision of the number of topics and concentration of information within topics.

3 Methods

We aim for scalability, thus we assume the existence of an unknown number of mixtures in any corpus. Nonparametric Bayesian methods are appropriate, as they define a model with an infinite limit of finite mixtures. In a hierarchical Dirichlet process (HDP) model, each data grouping is associated with a mixture model, which in turn contributes to a global model. Gibbs sampling is used for inference. Hyper parameters are determined by user-defined labels and word distributions within and across documents.

Our goal is to produce a semi-supervised topic model in which the topics align more or less tightly with pre-defined user categories. As part of this process, words in the vocabulary can have a more or less strong *a priori* association with these categories. Given a set of categories defined by the user, a parameter $\lambda \in [0, 1]$ controls the degree of supervision. When $\lambda = 0$, topics emerge from the data. When $\lambda = 1$, the initial topics will coincide with the labels in a one-to-one mapping.

In our framework, we assume that different metrics or procedures can be used to define the association between a user-defined category, such as an LCSH term, and words in the vocabulary. For this experiment, we associate words in the vocabulary to LCSH terms if they appear in the official definition of the term (see section 4).

In our semi-supervised hierarchical topic model, we incorporate knowledge from user-provided labels, together with the word distribution, to infer the model. There are two sets of parameters. The first set consists of the global level topic distribution over the corpus ($\boldsymbol{\theta}_0$ in $supervisedSampling$ algorithm), the local level topic distribution over documents ($\boldsymbol{\theta}_d$), and the word distribution over topics. The other is a set of hyper parameters such as the concentration parameters for global and local level topics (γ and α). We use blocked Gibbs sampling that infers the two sets of parameters in turn. More detailed explanations can be found in [10] and [5].

Our $supervisedHDP$ algorithm shown below starts with an initialization using user-defined labels, and uses this initial state to estimate initial hyper parameters. Lines 8 and 9 control the degree of supervision: $supLag$ sets which iterations to supervise, and λ controls the probability a word gets supervision. In our $supervisedSampling$

Algorithm 1: supervisedHDP

```
// initialization
for each document d ∈ D do
    for each word w ∈ V do
        supervisedSampling();
sample hyper parameters γ and α;
// main iteration
for iter < max iteration do
    for each document d ∈ D do
        for each word w ∈ V do
            if iter%supLag = 0 then
                X ∼ Bernoulli(λ);
                if X == 1 then
                    supervisedSampling();
                else
                    do ordinary HDP sampling;
            else
                do ordinary HDP sampling ;
    sample hyper parameters γ and α;
```

Algorithm 2: supervisedSampling

```
if wordLabelRelation contains w then
    sample a label l_{d,w} ∼ Mult(ψ_w);
    if labelToTopic has l_{d,w} then
        z_{d,w} ⟵ labelToTopic.get(l_{d,w});
    else
        z_{d,w} ⟵ a new topic;
        labelToTopic.put(l_{d,w}, z_{d,w});
else
    if no topic created in the corpus then
        t_{d,w} ⟵ a new table;
        z_{d,w} ⟵ a new topic;
        topicToTable_d.put(z_{d,w}, t_{d,w});
    else
        θ_0 ∼ Dir(γ);
        z_{d,w} ∼ Mult(θ_0);
if topicToTable_d has z_{d,w} then
    t_{d,w} ⟵ topicToTable_d.get(z_{d,w});
else
    θ_d ∼ Dir(α);
    t_{d,w} ∼ Mult(θ_d);
    topicToTable_d.put(z_{d,w}, t_{d,w});
```

algorithm, words first sample a label based on $wordLabelRelation$ (ψ_w), then local and global topics are inferred. $labelToTopic$ keeps track of the mapping between topics and labels, due to our assumption that a topic can be viewed as a mixture of labels (including none), and a label as a mixture of topics. They are different projections on different coordinates of interest.

4 Experiment

Data. The dataset contains 423 websites, addressing human rights issues all over the world. They are maintained by official organizations or individuals, and are in more than ten languages including English, Spanish, French, Chinese and Arabic. Each website home page and the depth one pages are crawled, concatenated and treated as a single document. We smooth our dataset to facilitate topic modeling. HTML headings and tags are removed. After we remove non-English websites, Flash sites and non-content ones, we end with a collection of 201 sites. We applied the Stanford CoreNLP named entity recognizer for organization, place and person names. The total size of the dataset is nearly 6 million words (30,000 per site). Removal of stop words and rare words, and concatenation of named entity words yields 29,392 unique terms.

Four Topic Models. We applied the following algorithms to the dataset: hierarchical Dirichlet process (HDP); supervision at initialization (InitialSup); supervision at initialization and every 20 iterations, with a 10% probability for words to be generated under supervision (LightlySup); supervision at every iteration, with a 50% probability for words to be generated under supervision (HeavilySup). For the three semi-supervised methods, librarians provided a set of labels (229 LCSH terms) related to human rights.

LCSH is a dynamic resource whose authoritative definitions are available on the web.[1] To assign word-level labels, we rely on the descriptive text that defines a term along with other textual fields (e.g., *General Notes*) and lists of variant terms, related terms and narrower terms. For each word w in our vocabulary, we associate w with each LCSH term l using the normalized term frequency of w in the descriptive text of the authority record for l. The word *police*, for example, is associated with four labels (LCSH terms) with the following strengths: *Bodyguards* (0.80), *Training* (0.09), *Peacekeeping forces* (0.07), *Extraordinary rendition* (0.04). About 10% of the vocabulary is associated with LCSH terms (2.9 per word on average).

User Study. We conducted a user study to measure how well users' assignments of topics to websites aligns with the relevant topic model under two conditions: the HDP and LightlySup topic models. For each condition, we randomly selected 15 sites, and from the topics for these sites, we selected 16 topics. Three librarians and three graduate students were recruited to participate in two one-hour sessions on different days to match topic word clouds to web sites. All did LightlySup on the first day, HDP on the second. Participants were instructed to browse each website and assign zero to three topics, along with percentages to reflect coverage.

5 Results

Two intuitive criteria that have been proposed for good topics are that they should be more fine-grained, and have fewer words in common [7]. Table 1 presents descriptive statistics for the results of the four methods at 200 iterations, illustrating that with greater supervision, topics are more numerous, and have fewer words in common. HeavilySup has 66.9% more topics than LightlySup, and 87.4% more than the average of HDP and InitialSup (col. 1). Columns two and three give the total number of distinct words among the top 1000 words across all topics for each method, and the distinct words per topic, followed by the ratio of topics per word. The first three methods have about the same ratio, while HeavilySup has half again as many.

To measure topic distinctness, we used the Jaccard coefficient [6], the ratio of the size of the intersection of two sets to the size of their union. Values range from 0 for disjoint sets to 1 for identical sets. Column 5 of Table 1 gives the average of the Jaccard coefficient for the sets of top twenty words from all pairs of topics for each method. The HDP topics have the most words in common, and LighlySup has the fewest.

Another contrast between the supervised and unsupervised methods pertains to coverage, in the sense of how many websites a topic gets assigned to (S/T, col. 6), and how much of a website it represents (T/S, col. 7). Supervised topics have higher S/T and

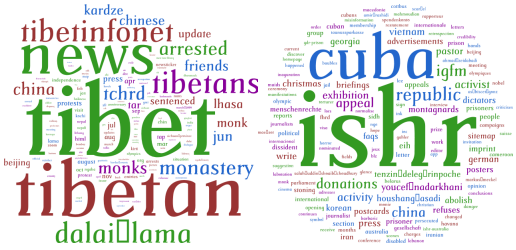
[1] `http://id.loc.gov/authorities/subjects.html`

lower Jaccard scores, which suggests they are relatively more distinct and less likely to be *vacuous* [1]. More supervised topics are assigned to each site (T/S) because each topic is more specific.

Table 1: Descriptive Statistics (200 Iterations). Jacc is Jaccard score of 10^{-3}; S/T refers to sites per topic and T/S refers to topics per site.

Method	Topics	Vocab.	Ratio	Jacc	S/T	T/S
HDP	131	22,233	0.0059	13.11	4.11	2.68
InitialSup	138	22,652	0.0061	10.78	6.07	4.16
LightlySup	151	24,259	0.0062	8.34	6.21	4.67
HeavilySup	252	26,740	0.0094	9.14	10.47	13.12

All the supervised methods are initialized by a set of topics corresponding directly to the human rights LCSH terms. Any new topic necessarily has no *a priori* association with an LCSH term, thus at any iteration after the first, some topics will link to an LCSH term and others might not. At the 200th iteration, 89% of LightlySup topics and 94% of HeavilySup topics are associated with LCSH terms. Figure 1 illustrates topics from HeavilySup with and without an associated LCSH term. Font size in the word cloud represents the probability of the word in the topic.



(a) *Government and the press*. (b) No LCSH term

Fig. 1: Two HeavilySup Topics

Figure 1a, associated with the LCSH term *Government and the press*, reflects characteristics specific to this dataset: websites related to this term are often about Tibet and freedom of the press in China. Here, supervision yields topics that relate the data to the terminology in ways that an unsupervised method would not. Figure 1b is a contrasting example of a topic that is not associated with an LCSH term. It accounts for a significant proportion of the content of two websites. Thus supervised topic modeling can still find hidden relations among sites.

5.1 User Study Results

Table 2: User results

	Topic	Prec	Rec	F
Librarians				
HDP	1.88	0.49	0.71	0.54
LightlySup	2.11	0.59	0.53	0.51
Students				
HDP	1.38	0.66	0.70	0.64
LightlySup	1.89	0.61	0.50	0.51
All				
HDP	1.63	0.58	0.70	0.59
LightlySup	2.00	0.60	0.51	0.51

On average, the LightlySup and HDP models assign 2.00 and 1.63 topics per site, respectively. As shown in col. 2 of Table 2, librarians assign topics at a more consistent rate across models (1.88 for HDP vs. 2.11 for LightlySup) while students' rates are less so (1.38 vs. 1.89). Columns 3-5 report precision/recall and f-measure (F) of the user assignments compared with the topic model assignments.

Precision on the user task is more critical to the librarians' ultimate goals than the other measures we report (recall and F), which are to have an automated method analogous to subject indexing. Librarians had much higher precision on the LightlySup task (0.59 vs. 0.49). This difference between conditions did not show up for students, who

had slightly higher precision on HDP (0.66 versus 0.61). A key difference between librarians and students was the much higher rate at which librarians assigned topics, suggesting a preference for finer-grained models, as has been reported elsewhere for subject matter experts [7]. Our design penalizes LightlySup in that by the time subjects did HDP they had had more practice on the task, and LightlySup assigned more topics per site (including two sites with 4 topics), so that probability of error was higher. Nevertheless, the overall results were roughly equivalent.

6 Conclusions

Supervised HDP yields topics that are more fine-grained, and more distinct, than topics produced by HDP. We tested the use of a controlled vocabulary in worldwide use by librarians for a subdomain pertaining to human rights, but the method can use any *a priori* semantic dimension. It is necessary to produce weighted (word,label) pairs. We tried various methods to do so, including mutual information between the subject terms used to catalog a website and words in the website text. This had noisy results due to a much larger number of (word,label) associations. In future work, we will compare different external resources, and continue to explore degrees of supervision.

Acknowledgements. We thank Terence H. Catapano, Melanie Wacker and Stuart Marquis for introducing us to the human rights archive, and we thank them and the additional librarians and students who participated in the study.

References

1. Alsumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic significance ranking of LDA generative models. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I. pp. 67–82. Berlin, Heidelberg (2009)
2. Andrzejewski, D., Zhu, X.: Latent dirichlet allocation with topic-in-set knowledge. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. pp. 43–48 (2009)
3. Blei, D.M., McAuliffe, J.D.: Supervised topic models. In: Advances in Neural Information Processing Systems (NIPS) (2007)
4. Blei, D.M., Ng, A., Jordan, M.: Latent dirichlet allocation. JMLR 3, 993–1022 (2003)
5. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588 (1995)
6. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles 44, 223–270 (1908)
7. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 262–272. Edinburgh, Scotland, UK. (July 2011)
8. Perotte, A., Bartlett, N., Elhadad, N., Wood, F.: Hierarchically supervised latent dirichlet allocation. In: Advances in Neural Information Processing Systems (NIPS) (2011)
9. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 248–256 (2009)
10. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581 (2006)