TSDPMM: Incorporating Prior Topic Knowledge into Dirichlet Process Mixture Models for Text Clustering

Linmei Hu[†], **Juanzi Li**[†], **Xiaoli Li**[‡], **Chao Shao**[†], **Xuzhong Wang**[§]
[†] Dept. of Computer Sci. and Tech., Tsinghua University, China
[‡] Institute for Infocomm Research(I2R), A*STAR, Singapore
[§] State Key Laboratory of Math. Eng. and Advanced Computing, China
{hulinmei1991, lijuanzi2008}@gmail.com
xlli@i2r.a-star.edu.sg, {birdlinux, koodoneko}@gmail.com

Abstract

Dirichlet process mixture model (DPMM) has great potential for detecting the underlying structure of data. Extensive studies have applied it for text clustering in terms of topics. However, due to the unsupervised nature, the topic clusters are always less satisfactory. Considering that people often have some prior knowledge about which potential topics should exist in given data, we aim to incorporate such knowledge into the DPMM to improve text clustering. We propose a novel model TSDPMM based on a new seeded Pólya urn scheme. Experimental results on document clustering across three datasets demonstrate our proposed TSDPMM significantly outperforms state-of-the-art DPMM model and can be applied in a lifelong learning framework.

1 Introduction

Dirichlet process mixture model (DPMM) (Neal, 2000) has been used in detecting the underlying structure in data. For example, (Vlachos et al., 2008; Vlachos et al., 2009) applied it to lexical-semantic verb clustering. (Wang et al., 2011; Huang et al., 2013; Yin and Wang, 2014) applied it for text clustering in terms of their topics. While DPMM achieved some promising results, it can still sometimes produce unsatisfactory topic clusters due to its unsupervised nature.

On the other hand, people often have prior knowledge about what potential topics should exist in a given text corpus. Take an earthquake event corpus as an example. The topics, such as "*casualties and damages*", "*rescue*" and "*government reaction*", called prior topics, are expected to occur in the corpus according to our common knowledge (e.g., the topics automatically learned from previous events using topic modeling (Ahmed and Xing, 2008)) or external resources (e.g., table of contents at Wikipedia event pages [1]). Similarly, in academic fields, "call for papers (CFP)" of conferences [2] lists main topics that conference organizers would like to focus on. Clearly, these prior topics can be represented as sets of words, which are available in many real-world applications. They can serve as weakly supervised information to enhance the unsupervised DPMM for text clustering.

Standard DPMM (Neal, 2000; Ranganathan, 2006) lacks a mechanism for incorporating prior knowledge. Some existing work (Vlachos et al., 2008; Vlachos et al., 2009) added knowledge of observed *instance*-level constraints (*must-links* and *cannot-links* between documents) to DPMM. (Ahmed and Xing, 2008) proposed recurrent Chinese Restaurant Process to incorporate *previous* documents with known topic clusters. We focus on incorporating *topic*-level knowledge, which is more challenging, as seed/prior topics could be latent rather than observable.

Particularly, we construct our novel TSDPMM (Topic Seeded DPMM) based on a principled seeded Pólya urn (sPU) scheme. Our model inherits the nonparametric property of DPMM and has additional technical merits. Importantly, our model is encouraged but not forced to find evidences of seed topics. Therefore, it has freedom to discover new topics beyond prior topics, as well as to detect which prior topics are not covered by current data. It is thus convenient to observe topic variations between prior topics and newly mined topics. Experimental results on document clustering across three corpora demonstrate that our model effectively incorporates prior topics, and significantly outperforms state-of-the-art DPMM model. Particularly, our TSDPMM can be applied in a lifelong learning framework which enables the prior

[1] e.g., http://en.wikipedia.org/wiki/2010_Chile_earthquake

[2] e.g., https://nips.cc/Conferences/2014/CallForPapers

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 787–792,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

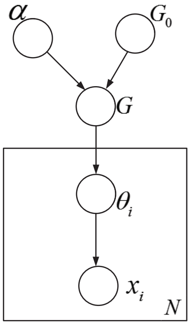
topic knowledge to evolve as more and more data are observed.

2 Topic Seeded DPMM

In this section, we first introduce the standard DPMM model for document clustering in terms of topics. Then we describe how to incorporate seed/prior topics into the model using a seeded Pólya urn (sPU) scheme, which gives us our novel TSDPMM model (Topic Seeded DPMM). Finally, we present the model inference.

2.1 DPMM

The DPMM (Antoniak, 1974) as a non-parametric model assumes the given data is governed by an infinite number of components where only a fraction of these components are activated by the data. Figure 1 illustrates the DPMM graphical model and its generative process of a document x_i. First, we sample a topic $\theta_i = \{\theta_{ij}\}_{j=1}^{j=|V|}$ (a multinomial distribution over words belonging to the vocabulary V) for the document x_i according to a Dirichlet Process (DP) $G \sim DP(\alpha, G_0)$, where $\alpha > 0$ is a concentration parameter and the base measure $G_0 = Dir(\vec{\beta})$ can be considered as a prior distribution for θ. Consider the document x_i as a bag of words, given the topic θ_i, the generative distribution F is a given likelihood function parameterized by θ. We define F as $p(x_i|\theta_i) = \prod_{j=1}^{|x_i|} p(x_{ij}|\theta_i)$, where x_{ij} is the j_{th} word in x_i. Note that the DPMM assumes each document can be assigned to one topic cluster only.



$$G \sim DP(\alpha, G_0)$$
$$\theta_i|G \sim G$$
$$x_i|\theta_i \sim F(\theta_i)$$

Figure 1: Graphical Representation of DPMM.

The DP process of DPMM, according to which topic θ_i for a document x_i is drawn, can be explained by the popular metaphor of Pólya urn (PU) scheme (Blackwell and MacQueen, 1973), equivalent to the Chinese Restaurant Process (Ahmed and Xing, 2008). The PU scheme works on balls (documents) and colors (topics). It starts with an empty urn. With probability proportional to α, we draw $\theta_i \sim G_0$, and add a ball of this color to the urn. With probability proportional to $i - 1$ (i.e., the current number of balls in the urn), we draw a ball at random from the urn, observe its color θ_i and replace the ball with two balls of the same color. In this way, we draw topic θ_i for document x_i. As shown in the process, the prior probability of assigning a document to a topic is proportional to the number of documents already assigned to the topic. As a result, the DPMM exhibits the "rich get richer" property.

2.2 TSDPMM: Incorporating Seed Topics

In this section, we describe our proposed algorithm to incorporate prior seed topics into the DPMM. A prior/seed topic k is represented by a vector $\vec{N}_k^{(0)}$ (word frequencies under the topic). We can obtain the prior topics represented by $\vec{N}_k^{(0)}$ from past learning of topic models or external resources such as Wikipedia and "CFP". Assuming we have $K^{(0)}$ prior topics, we use the parameter $\vec{\alpha}^{(0)} = \{\alpha_k^{(0)}\}_{k=1}^{K^{(0)}}$ to control our confidence about how likely each prior topic exists. Let us go back to Pólya urn (PU) scheme, where a prior topic can be taken as a known color. We extend the PU scheme to incorporate prior topics, which gives the **sPU** (seeded Pólya Urn) scheme. The sPU scheme can be described as follows:

- We start with an urn with $\alpha_k^{(0)}$ balls of each known color $k \in \{1, ..., K^{(0)}\}$.
- With a probability proportional to α, we draw $\theta_i \sim G_0$ and add a ball of this color to the urn.
- With probability proportional to $i - 1 + \sum_{k=1}^{K^{(0)}} \alpha_k^{(0)}$, we draw a random ball from the urn, and replace the ball with two balls of the same color.

As shown in the above process, instead of starting with an empty urn in DPMM, we assume that the urn already has certain balls of known colors. In this way, we incorporate the prior seed topics. The number of initial balls (documents) $\alpha_k^{(0)}$ controls how likely the topic k exists. We can use different values of $\alpha_k^{(0)}$ for prior topics with different confidence levels. This sPU scheme gives our novel model TSDPMM (Topic Seeded DPMM) incorporating prior topics. The TSDPMM has similar graphical representation as DPMM (Figure 1), except the introduction of hyper-parameter $\vec{\alpha}^{(0)}$. We then present a collapsed gibbs sampling algorithm for model inference as follows.

TSDPMM Inference. The model inference is described in detail in Algorithm 1. It first initializes all documents with random topic clusters. Then it iteratively updates the topic cluster assignments of documents according to the conditional probabilities (Eq.1) until convergence. Eq.1 can be derived as:

$$p(z_i|\vec{Z}_{-i},\vec{X}) \propto p(z_i|\vec{Z}_{-i},\alpha,\vec{\alpha}^{(0)})p(x_i|\vec{X}_{-i},\vec{Z},\vec{\beta}) \tag{1}$$

where z_i is the topic assignment of observation x_i, $\vec{X}$ is the given document corpus, and $\vec{Z}_{-i}$ are $\vec{X}_{-i}$ are the set of topic assignments and the corpus excluding the i_{th} observation x_i, respectively.

Algorithm 1: Collapsed Gibbs Sampling

```
Input: Document dataset $\vec{X} = \{x_i\}_{i=1}^{m}$, prior
       topics $\{\vec{N}_k^{(0)}\}_{k=1}^{K^{(0)}}$, parameter $\vec{\alpha}^{(0)}$
Output: Topic assignments $\vec{Z}$ of all
        documents
Initialize the topic assignments $\vec{Z}$ based on
prior topics randomly;
repeat
    Select a document $x_i \in \vec{X}$ randomly
    Fix the other topic assignments $\vec{Z}_{-i}$
    Assign a new value to
    $z_i$:$z_i \sim p(z_i|\vec{Z}_{-i},\vec{X})$(Eq. 1)
until Convergence;
```

In Eq.1, the first item $p(z_i{=}k|\vec{Z}_{-i},\alpha,\vec{\alpha}^{(0)})$ denotes a prior probability of $z_i{=}k$, which is proportional to the number of documents already assigned to it. If k is a **prior** topic, it is proportional to $n_{k,-i}+\alpha_k^{(0)}$, where $n_{k,-i}$ is the number of documents of topic k excluding the current document x_i. If k is an **existing** (not prior) topic, it is proportional to $n_{k,-i}$. If k is a **new** topic, the probability is proportional to α. The second item $p(x_i|\vec{X}_{-i},\vec{Z}_{-i},z_i = k,\vec{\beta})$ is the likelihood of x_i given $\vec{X}_{-i}$, $\vec{Z}_{-i}$ and $z_i{=}k$. They can be derived as $p(x_i|\vec{X}_{-i},\vec{Z}_{-i},z_i = k,\vec{\beta}) \propto \frac{p(\vec{X}|\vec{Z},\vec{\beta})}{p(\vec{X}_{-i}|\vec{Z}_{-i},\vec{\beta})}$ where $p(\vec{X}|\vec{Z},\vec{\beta}) = \int p(\vec{X}|\vec{Z},\Theta)p(\Theta|\vec{\beta})d\Theta$. As $p(\Theta|\vec{\beta})$ is a Dirichlet distribution and $p(\vec{X}|\vec{Z},\Theta)$ is a multinomial distribution, we can get $p(\vec{X}|\vec{Z}){=}\prod_{k=1}^{K}\frac{\Delta(\vec{N}_k+\vec{\beta})}{\Delta(\vec{\beta})}$, where $\vec{N}_k = \{N_{k,w}\}_{w=1}^{V}$ and $N_{k,w}$ is the number of occurrences of word w in the k_{th} topic. Here, we adopt the function Δ in (Heinrich, 2009), and we have $\Delta(\vec{\beta}) = \frac{\prod_{w=1}^{V}\Gamma(\beta)}{\Gamma(\sum_{w=1}^{V})\beta}$ and $\Delta(\vec{N}_k+\vec{\beta}) = \frac{\prod_{w=1}^{V}\Gamma(N_{k,w}+\beta)}{\Gamma(\sum_{w=1}^{V}(N_{k,w}+\beta))}$. Finally, we can derive:

$$p(z_i = k|\vec{Z}_{-i},\vec{X}) \propto \begin{cases} (n_{k,-i}+\alpha^{(0)})\cdot\frac{\Delta(\vec{N}_{.,i}+\vec{N}_{k,-i}+\vec{N}_k^{(0)}+\vec{\beta})}{\Delta(\vec{N}_{k,-i}+\vec{N}_k^{(0)}+\vec{\beta})} & \text{prior} \\ n_{k,-i}\cdot\frac{\Delta(\vec{N}_{.,i}+\vec{N}_{k,-i}+\vec{\beta})}{\Delta(\vec{N}_{k,-i}+\vec{\beta})} & \text{existing} \\ \alpha\cdot\frac{\Delta(\vec{N}_{.,i}+\vec{\beta})}{\Delta(\vec{\beta})} & \text{new ,} \end{cases}$$

where $\vec{N}_{k,-i}$ is a vector with the word counts for all the documents assigned to topic k excluding x_i, $\vec{N}_{.,i}$ and $\vec{N}_k^{(0)}$ are vectors with word counts in document x_i and in all the documents assigned to k in prior knowledge respectively. According to this equation, documents are likely to go into clusters which are bigger and give higher likelihood of the documents. When the Gibbs sampler converges, we obtain topic cluster assignments of all the documents. Different from DPMM inference process in which topics are removed when no documents is assigned to them, TSDPMM inference can retain prior topics all the time due to the initial number of documents $\vec{\alpha}^{(0)}$, making it able to track prior topics, as well as to detect new topics.

3 Experiments

We evaluate our proposed TSDPMM model for document clustering on 3 datasets where each cluster corresponds to a topic. We implement both DPMM and TSDPMM models — their source codes are available at `https://github.com/newsminer/DPMM_and_TSDPMM`.

3.1 Datasets

We collect machine learning conference NIPS datasets composed of paper titles and abstracts from 2012 to 2014 – each year includes 342, 360 and 411 documents respectively. They are named as NIPS-12, NIPS-13 and NIPS-14.

We also employ the standard *benchmark* news datasets, including 20 Newsgroups [3] and Reuters-21578. As news is often timely reported, we choose three continuous days with the largest number of documents in 20 Newsgroups (i.e. 11, 12 and 13 May) and Reuters-21578 (i.e. 3, 4 and 5 March) for our experiments. These datasets are

[3] http://people.csail.mit.edu/jrennie/20Newsgroups/

denoted as 20N-1, 20N-2, 20N-3 (including 103, 96, 106 documents) and Reu-1, Reu-2, Reu-3 (including 282, 249, 207 documents), respectively.

For all the datasets, we conduct the following preprocessing: (1) Convert letters into lowercase; (2) Remove non-Latin characters and stop words; (3) Remove words with *document frequency* < 2.

3.2 Experimental Setup

We take the standard DPMM as our baseline method and compare it with our proposed TSDPMM model using *different* prior knowledge obtained with different manners.

For NIPS datasets, we use two kinds of prior knowledge: one is the topics learned by DPMM from *previous* year's dataset; the other one is from an *external* resource "CFP" [4] (10 topics, same for each year). We name them as TSDPMM-P and TSDPMM-E respectively. As the topic descriptions in "CFP" are sparse, we repeat each topic description by ten times and then represent a topic with the words with word frequencies in its description text.

For both 20 Newsgroups and Reuters datasets, we use prior knowledge learned by DPMM from the previous day's dataset. Furthermore, to test if we can improve the results *continuously* by applying TSDPMM, every time when we model a new dataset, we incorporate prior topics learned by TSDPMM from previous day's dataset, similar to lifelong learning (Chen and Liu, 2014; Thrun, 1998). We call this model as TSDPMM-L.

Parameter Setting. Following a previous work (Vlachos et al., 2009), we set the hyper-parameters α=1, $\vec{\alpha}^{(0)}$=$\{1.0\}$, $\vec{\beta}$ =$\{1.0\}$. We run Gibbs sampler for 100 iterations and stop the iteration once the log-likelihood of the training data converges.

Evaluation. The widely used NMI (normalized mutual information) measure (Dom, 2002), has been employed to evaluate document clustering results. The higher a value of NMI, the better a clustering result is. However, NMI needs true class labels for documents, and can only be applied to our benchmark news datasets. For NIPS datasets without true labels, we use the measure of *perplexity*, as defined in (Blei et al., 2003), to test per-word likelihood of the datasets. The lower the perplexity, the better a model fits the data.

3.3 Results

Table 1 shows the average perplexity values of five runs of 3 models on NIPS datasets. It shows that both TSDPMM-P and TSDPMM-E, leveraging prior topics from previous learning and "CFP" significantly outperform DPMM. In addition, TSDPMM-E achieves lower performance than TSDPMM-P due to its lower quality of prior topics directly obtained from "CFP", compared to higher quality topics from past learning. We may improve "CFP" knowledge by extending it with related texts from search engines or Wikipedia using keywords in "CFP" in future work.

An insight of our clustering results on NIPS-14 dataset suggests that most prior topics in 2013 are covered again in 2014 (consistent topics), except a few missing topics such as "*lasso for Bayesian networks*". Additionally, some newly evolved topics in 2014, e.g. "*monte carlo particle filtering*" and "*nash games*", are successfully discovered by our proposed model.

Models	NIPS-12	NIPS-13	NIPS-14
DPMM	321.7	317.1	362.9
TSDPMM-P	**290.1**	**298.7**	**346.8**
TSDPMM-E	307.5	315.6	360.1

Table 1: Average perplexity of different models on NIPS.

Table 2 illustrates the average NMI values of five runs of DPMM, TSDPMM and TSDPMM-L on news datasets. The results show that TSDPMM using prior topics learnt by DPMM outperforms DPMM (on average **+5.8%**; p <0.025 with *t-test*). Additionally, TSDPMM-L, which continuously uses prior topics learnt by TSDPMM from previous dataset, further outperforms TSDPMM (on average **+3.2%**; p <0.025 with *t-test*). Note TSDPMM-L uses TSDPMM results of 20N-1 and Reu-1 as prior knowledge for the first time, so there are no TSDPMM-L results for the first days in Table 2 for 20N-1 and Reu-1 respectively.

3.4 Discussion

The experimental results across 3 datasets have demonstrated that our proposed models can improve DPMM model by incorporating prior topic knowledge, and the higher-quality knowledge will lead to better results. By applying our TSDPMM in a lifelong continuous learning framework, namely TSDPMM-L, can further improve

[4] https://nips.cc/Conferences/2014/CallForPapers

Models	20N-1	20N-2	20N-3	Reu-1	Reu-2	Reus-3
DPMM	0.610	0.537	0.590	0.509	0.647	0.653
TSDPMM	0.645	0.610	0.681	0.648	0.654	0.655
TSDPMM-L	–	**0.681**	**0.697**	–	**0.689**	**0.656**

Table 2: Average NMI of different models on news datasets.

text clustering due to the better prior topic knowledge obtained in the evolving environment.

4 Related Work

Our work is related to papers (Vlachos et al., 2008; Vlachos et al., 2009), which added supervision (*instance*-level must-links or cannot-links between documents) to the DPMM. (Ahmed and Xing, 2008) proposed recurrent Chinese Restaurant Process to incorporate previous documents with known topic clusters. However, our work is very different as we focus on how to incorporate latent *topic*-level prior knowledge. We model prior topics as known colors that have a certain probability proportional to $\alpha_k^{(0)}$ to be assigned to a document. In addition, our inference mechanism subsequently takes the prior knowledge into consideration for automatically assigning topics to documents.

Some existing studies such as (Ramage et al., 2009; Andrzejewski et al., 2009; Jagarlamudi et al., 2012; Andrzejewski et al., 2011) worked on incorporating prior lexical or domain knowledge into LDA. Different from all these work, we focus on the nonparametric model DPMM and propose to incorporate the prior topic knowledge obtained in multiple ways.

5 Conclusion

In this paper, we propose a novel problem of incorporating prior topics into DPMM model and address it through a simple yet principled seeded Pólya urn scheme. We show that the topic knowledge can be obtained in multiple ways. Experiments on document clustering across 3 datasets demonstrate our proposed model can effectively incorporate the prior topic knowledge and significantly enhance the standard DPMM for text clustering. In future work, we will study how to discover overlapping clusters, i.e., allowing one document to be grouped into multiple topic clusters. We will also explore how to incorporate prior knowledge about topic relations (such as causation and correlation) into topic modeling.

Acknowledgments

The work is supported by 973 Program (No. 2014CB340504), NSFC-ANR (No. 61261130588), Tsinghua University Initiative Scientific Research Program (No. 20131089256), Science and Technology Support Program (No. 2014BAK04B00), and THU-NUS NExT Co-Lab.

References

Amr Ahmed and Eric P Xing. 2008. Dynamic nonparametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *SDM*, pages 219–230. SIAM.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual ICML*, pages 25–32. ACM.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *Proceedings-IJCAI*, volume 22, page 1171.

Charles E Antoniak. 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

David Blackwell and James B MacQueen. 1973. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355.

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

Zhiyuan Chen and Bing Liu. 2014. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference*, pages 1116–1125. ACM.

Byron E Dom. 2002. An information-theoretic external cluster-validity measure. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 137–145. Morgan Kaufmann Publishers Inc.

Gregor Heinrich. 2009. Parameter estimation for text analysis. Technical report, vsonix GmbH and University of Leipzig.

Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi. 2013. Dirichlet process mixture model for document clustering with feature partition. *Knowledge and Data Engineering,*, 25(8):1748–1759.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the ACL*, pages 204–213. Association for Computational Linguistics.

Radford M Neal. 2000. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Daniel Ramage, David Leo Wright Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on EMNLP*, pages 248–256.

Ananth Ranganathan. 2006. The dirichlet process mixture (dpm) model. Technical report, Citeseer.

Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer.

Andreas Vlachos, Zoubin Ghahramani, and Anna Korhonen. 2008. Dirichlet process mixture models for verb clustering. In *Proceedings of the ICML workshop on Prior Knowledge for Text and Language*.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 74–82. Association for Computational Linguistics.

Chan Wang, Caixia Yuan, Xiaojie Wang, and Wenwei Xue. 2011. Dirichlet process mixture models based topic identification for short text streams. In *NLP-KE*, pages 80–87. IEEE.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD*, pages 233–242. ACM.