2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Topic Discovery from Heterogeneous Texts

Jipeng Qiang[1,2], Ping Chen[2], Wei Ding[2], Tong Wang[2], Fei Xie[1,3], Xindong Wu[1,4]
[1]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China*
[2]*Department of Computer Science, University of Massachusetts Boston, Boston, MA 02155*
[3]*Department of Computer Science and Technology, Hefei Normal University, Hefei 230009, China*
[4]*School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, Louisiana 70504*
Email: qjp2100@gmail.com

***Abstract*—Recently many topic models such as Latent Dirichlet Allocation (LDA) have made important progress towards generating high-level knowledge from a large corpus. They assume that a text consists of a mixture of topics, which is usually the case for regular articles but may not hold for a short text that usually contains only one topic. In practice, a corpus may include both short texts and long texts, in this case neither methods developed for only long texts nor methods for only short texts can generate satisfying results. In this paper, we present an innovative method to discover latent topics from a heterogeneous corpus including both long and short texts. A new topic model based on collapsed Gibbs sampling algorithm is developed for modeling such heterogeneous texts. The experiments on real-world datasets validate the effectiveness of the proposed model in comparison with other state-of-the-art models.**

***Index Terms*—Topic Model, LDA, Heterogeneous texts, Collapsed Gibbs Sampling**

1. Introduction

With the rapid development of the World Wide Web, the number of electronic texts is huge. Therefore, such data cannot be handled by human power and hence demands effective automated tools. Topic models have been proven to be useful for automatic topic discovery from a huge volume of texts. A topic model views texts as a mixture of probabilistic topics, where a topic is represented by a probability distribution over words [1]. Based on the assumption that each text of a collection is modeled over a set of topics, many topic models such as Latent Dirichlet Allocation (LDA) have demonstrated great success on long texts [2], [3], [4], [5].

Along with the emergence and popularity of social media (e.g. Twitter and Facebook), short texts also have been an important information source. Because LDA cannot work very well on short texts, how to extract topics from short texts becomes a research problem [6], [7], [8]. Compared with long texts, short texts usually contain much less information. Therefore, it is challenging to discover topics for short texts, due to the fact that only very limited word co-occurrence information is available in short texts compared with long texts. The simple assumption that each text is sampled from only one latent topic is totally unsuited to long texts, but it can be suitable for short texts [9], [10]. Therefore, many models for short texts were proposed based on this simple assumption [10], [11], [12] (e.g. DMM [12] and BTM [11]). The other strategy takes advantages of various heuristic ties among short texts to aggregate them into long pseudo-texts before topic inference [6], [13], [14], [15]. However, these schemes are heuristic and highly dependent on the data. Furthermore, such metadata may not be available for short texts such as news titles, advertisements or image captions.

In reality, we often collect text data from different sources (e.g. news websites, Facebook, Twitter, etc.) for topic discovery without regard to the length of a text. Here, news articles are often long. Tweets and news comments can be short texts. In this paper, the corpus includes both short texts and long texts, referred to as heterogeneous texts. Table 1 shows the results on heterogeneous texts including four hot issues (”Oculus Rift”, ”SpaceX Rocket”, ”Donald Trump”, and ”Windows 10”) on January 6, 2016, which are collected from different news websites and Twitter. How to discover latent topics from heterogeneous texts? There are several native approaches as follows to solve this problem based on the current models.

In the first approach, we adopt long text topic modeling on only long texts after removing short texts, or we adopt short text topic modeling on only short texts after removing long texts. Although a short text generally contains less information than a long text, a corpus often includes a large number of short texts. Therefore, it is not reasonable if we remove short texts or long texts from the corpus. From Table 1, we can see that LDA on only long texts cannot identify the topic ”SpaceX Rocket”, and DMM on only short texts only identifies one topic ”Oculus Rift”. Another approach is that we directly adopt long text topic modeling (e.g. LDA [3]) on the whole corpus, or we directly adopt short text topic modeling (e.g. DMM [12] or BTM [11]). If we only use the assumption that each text is sampled from one topic, the model will have poor performance for long texts. On the other hand, if the assumption that each text is modeled over multiple topics, the model probably cannot work well for short texts. Table 1 shows that LDA

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.36

IEEE computer society

and BTM on the whole data cannot identify the two topics, "Oculus Rift" and "Windows 10", and DMM on the whole data cannot identify the two topics ("SpaceX Rocket" and "Donald Trump"). In the last approach, we adopt short text topic modeling for short texts and long text topic modeling for long texts, and then aggregate the results. The word co-occurrence information of the whole data is divided into two parts that will result in two worse results learned by each model. In addition, how to merge the two results remains a hard problem, for example, from DMM(short texts) and LDA(long texts).

With the above analysis, we formulate a novel problem on how to extract topics from heterogeneous texts. It is natural for us to consider the two assumptions simultaneously, namely the simple assumption about sampling a topic for short text and the complex assumption about sampling multiple topics for long text, respectively. Therefore, we design a new Heterogeneous Text Topic Modeling (HTTM) to discover topics from heterogeneous texts by taking advantage of both assumptions. From Table 1, we can see that the topics learned by our method HTTM are largely better than those learned by the native methods. Our contributions are summarized as follows.

- This is the first attempt to discover latent topics from heterogeneous texts.
- We propose Heterogeneous Text Topic Modeling (HTTM), a collapsed Gibbs Sampling algorithm for topic discovery to handle heterogeneous texts by sampling a topic for each short text and multiple topics for each long text of a collection, unlike previous work for any type of texts under a single assumption.
- We evaluate our model on real-world datasets, and the experimental results demonstrate that our model achieves substantial improvement over the state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents our topic model from heterogeneous texts. Section 4 shows experimental results. Finally, Section 5 concludes the paper.

2. Related Work

Nigam et al. [16] proposed a mixture of unigram model based on the assumption that each document is generated by one topic. This simple assumption is often too limited to effectively model a large collection of long texts. The complex assumption that each text is modeled over multiple topics was widely used by topic discovery from long texts [4], [17]. In a sense, the complex assumption does capture the possibility that a document may contain multiple topics. Based on this assumption, many topic models such as Latent Dirichlet Allocation (LDA) have shown promising results [2]. Variational Bayes [2] and collapsed Gibbs Sampling [3] are two commonly used approximate inference methods for learning LDA and other models, such as author-topic models [18] and relational topic models [11].

As a lot of short texts have been collected from social networks such as Twitter, many people analyze this type of data to find latent topics using different topic models [9], [19], [20]. Many topic models such as LDA have not been able to work very well on short texts because only very limited word co-occurrence information is available in short texts [14]. Meanwhile, the assumption that each text is generated by one topic does not fit long texts. A lot of research has shown that this assumption works well on short texts [7], [12]. Therefore, many topic models adopted this assumption for topic discovery in short texts. Zhao et al. [10] empirically compared the data with traditional news media, and proposed a Twitter-LDA model by assuming that one tweet is generated from one topic. Yin and Wang [12] also adopted this assumption for topic inference based on Gibbs Sampling. The other strategy takes advantages of various heuristic ties among short texts to aggregate them into long pseudo-texts before topic inference [13], [14]. In a sense, each short text is considered to be generated from a long pseudo-text. The strategy can be regarded as an application of the author-topic model [18] to tweets, where each tweet (text) has a single author. For example, some models aggregated all the tweets of a user as a single text [6], [15]. As these tweets with the same hashtag may come from a topic, Mehrotra et al. [13] aggregated all tweets into a pseudo-text based on hashtags.

In this paper, we want to discover latent topics from heterogeneous texts which include both short texts and long texts. Although there are many existing works on topic inference for short texts or long texts, there is no previous work for topic discovery from heterogeneous texts. For heterogeneous texts, any one of the two assumptions separately is adopted that may lead to poor inference. For example, it is unreasonable to model a long text to contain only one topic, or model a short text to contain multiple topics. Motivated by this, a natural strategy is to incorporate the two assumptions together into topic inference. To the best of our knowledge, the proposed topic model is the first one focusing on heterogeneous texts, which does not exploit any external knowledge.

3. Topic Modeling from Heterogeneous Texts

In this section, we present a novel topic model (HTTM, Heterogeneous Text Topic Modeling) by integrating the two assumptions that each long text is represented as the mixing proportions of multiple latent topics and each short text only contains one latent topic.

3.1. Model Description

The generative process of HTTM from heterogeneous texts is shown in Figure 1. Let W represent a vocabulary, S a collection of short texts, L a collection of long texts, and D the whole set. The rightmost part in Figure 1 follows the assumptions of standard topic models (e.g. LDA) to generate a set of long texts L. The leftmost part is used to generate a set of short texts S, which corresponds to a mixture of

TABLE 1. TOPICS LEARNED FROM NEWS&TWEET DATASET. TOP-10 WORDS FROM FOUR TOPICS COMPARED WITH FIVE NATIVE METHODS. WORDS THAT ARE NOISY AND LACK OF REPRESENTATIVENESS ARE HIGHLIGHTED IN BOLD

Topic	Method	Top Words
Oculus	LDA	**windows oculus microsoft rift devices company headset january lumia hours**
	DMM	oculus rift **video** open headset reality free preorders **launch** virtual
	BTM	**windows microsoft oculus rift devices company news headset hours lumia**
	DMM(Short Texts)	oculus rift **video** open headset reality preorders free virtual **launch**
Rift	LDA(Long Texts)	oculus rift headset january virtual reality company wednesday touch price
	HTTM	oculus rift headset virtual reality january wednesday company touch price
SpaceX	LDA	rocket spacex falcon space landing launch musk stage **company** satellites
	DMM	**windows cosby year trump microsoft rocket china monday spacex time**
	BTM	rocket spacex falcon space launch landing musk stage **company** satellites
	DMM(Short Texts)	**coughlin strickland craig windows chip kelly fired tablet better donald**
Rocket	LDA(Long Texts)	**cosby rocket spacex falcon space landing launch bill musk women**
	HTTM	rocket spacex falcon space landing launch musk stage **company** satellites
Donald	LDA	trump campaign donald iowa cruz republican clinton state hampshire states
	DMM	**coughlin bill cosby giants trump donald odell beckham coach video**
	BTM	trump campaign iowa cruz republican donald hampshire clinton rubio states
	DMM(Short Texts)	**trump donald video recruitment featured affiliate qaeda terror comments group**
Trump	LDA(Long Texts)	trump campaign iowa cruz republican donald state hampshire clinton states
	HTTM	trump campaign donald iowa cruz republican clinton state hampshire states
Windows	LDA	**email comments post news today comment account facebook access badge**
	DMM	windows microsoft **keys** mobile **encryption** tablet users **acer liquid jade**
	BTM	**died year attack saudi january oregon fire iran star best**
	DMM(Short Texts)	**windows china microsoft stocks mobile food keys xbox encryption users**
10	LDA(Long Texts)	windows microsoft devices lumia company percent mobile running **hours** system
	HTTM	windows microsoft devices lumia company mobile percent users running december

unigrams that each text is sampled from one latent topic. The above generative procedure can be described in Algorithm 1.

Algorithm 1: HTTM

1: **for** all topics $k \in [1,K]$ **do**
2: sample a word distribution $\phi_k \sim \text{Dir}(\beta)$
3: **end for**
4: **for** all texts $d \in [1,D]$ **do**
5: choose $N \sim \text{Poisson}(\xi)$
6: **if** $N >$ length limit **then** // long text
7: draw $\theta \sim \text{Dir}(\alpha)$
8: **for** each of the N words **do**
9: draw a topic $z_{li} \sim \text{Multi}(\theta)$
10: $w \sim \text{Multi}(\phi_{z_{li}})$
11: **end for**
12: **else** // short text
13: draw a topic z_s from $[1,K]$
14: **for** each of the N words **do**
15: $w \sim \text{Multi}(\phi_{z_s})$
16: **end for**
17: **end if**
18: **end for**

Here, α and β are hyperparameters of the Dirichlet priors, $\phi_{z_{li}}$ refers to the multinomial distribution over words for topic z_{li}, and all words of a short text are sampled from a topic ϕ_{z_s}. Steps 7 to 11 of the algorithm corresponds to the process of generating long texts, and Steps 13 to 16 corresponds to the process of generating short texts.

Long text topic models estimate hidden variables z_l, and short text topic models estimate hidden variable z_s. The key problem of HTTM is to estimate the posterior

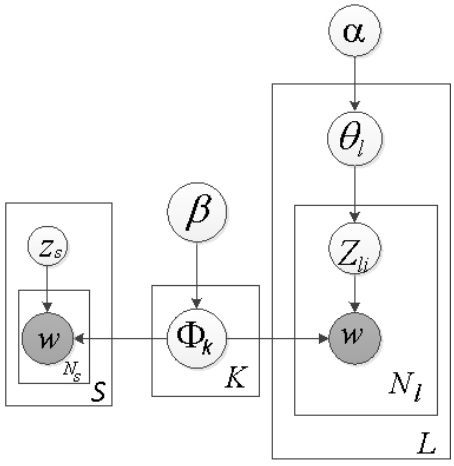


Figure 1. Graphical representation of the Heterogeneous Texts Topics Modeling. α and β are hyperparameters, z_s is the hidden document topic for generating the s^{th} short text of S, z^{li} denotes the topic identities assigned to the i^{th} word w_{li} in the l^{th} long text of L, ϕ represents the distribution of words in topics, and θ is the topic distribution for long texts.

distribution of the hidden variables z_s and z_l for a given piece of heterogeneous text simultaneously, p(z_s, z_l, ϕ, θ | α, β). Below, we will use collapsed Gibbs Sampling to estimate parameters under Dirichlet priors.

Using the collapsed Gibbs Sampling for topic discovery from heterogeneous texts, we need to deal with two sets of latent variables (z_l and z_s) separately, and update the parameters (θ and ϕ) in a unified framework.

When dealing with long texts, the multinomial distributions of ϕ and θ can be integrated, and the conditional

distribution is given by,

$$p(z_{li} = k \mid D^{-li}, \alpha, \beta) \propto \frac{(n_l^{k,-li} + \alpha)(n_k^{w_{li},-li} + \beta)}{n_k^{-li} + V\beta} \quad (1)$$

where z_{li} refers to topic *k* of the i^{th} word w_{li} in the l^{th} text of *L*, $p(z_{li} \mid D^{-li}, \alpha, \beta)$ is the probability of word w_{li} belonging to topic *k* conditioned on the whole set *D* after removing the current word, n_l^k is the number of occurrences of topic *k* in text *l*, $n_k^{w_{li}}$ is the number of occurrences of word w_{li} belonging to topic *k* in *D*, n_k is the number of occurrences of all words *W* belonging to topic *k* in *D*, and *V* is the size of *W*. Moreover, the superscript $.^{-li}$ means topic *k* of word w_{li} in text *l* is removed from z_l. Here, *K* is the number of topics in *D*, and *k*=1,2,...,*K*. For instance, n_k^{-li} is obtained through removing the (topic, word) combination at the i^{th} word of the l^{th} text. If all texts in the dataset are long texts, HTTM is reduced to LDA.

For short texts, we can infer the latent variable z_ss the conditional distribution $p(z_s = k|D^{-s})$ as follows,

$$p(z_s = k \mid D^{-s}, \beta) \propto \frac{\prod_{w \in d_s}(n_k^{w,-s} + \beta)}{\prod_{i=1}^{N_s}(n_k^{-s} + V\beta)} \quad (2)$$

where the superscript $.^{-s}$ means the s^{th} short text is excluded from D, $n_k^{w,-s}$ is the number of occurrences of word w belonging to topic *k* without considering the s^{th} short text . Here, HTTM is different from short text topic models Unigrams [16] and DMM [12] about inferring the latent variable z_s. z_s in Unigrams and DMM tends to choose a topic with more texts. Because a long text contains multiple topics, we only consider the impact of words in short texts, and the detailed derivation is shown below in Section 3.2.

Finally, using the counts of topic assignments of long texts and short texts, we can estimate the topic-word distribution of ϕ and text-topic distribution θ as follows,

$$\phi_k^w = \frac{n_k^w + \beta}{n_k + V\beta}$$

$$\theta_k^l = \frac{n_k^l + \alpha}{n_l + K\alpha}$$

$$\theta_k^s = \begin{cases} 1, k = z_s \\ 0, others \end{cases}$$

where ϕ_k^w denotes that the probability of word *w* is generated by topic *k*, and can be regarded as the importance of word *w* to topic *k*, n_l is the number of words in long text l, and n_k^l is the number of words belonging to topic *k* in text l.

3.2. Parameter Estimation

For short texts, we need to infer the latent variable z_s based on collapsed Gibbs Sampling. The conditional distribution $p(z_s = k \mid D^{-s})$ is as follows:

$$p(z_s = k \mid D^{-s}) \propto \frac{p(z_s \mid D, \beta)}{p(z_s \mid D^{-s}, \beta)}$$

where $p(z_s = k \mid D^{-s})$ is the probability of text d_s belonging to topic *k*, and $.^{-s}$ means that the s^{th} short text is removed from *D* and z_s. For obtaining $p(z_s = k \mid D^{-s})$, we can marginalize out the random variable ϕ, and get $p(z_s = k \mid D^{-s}) = \int p(z_s \mid \phi, D)p(\phi \mid \beta)d\beta$. Here, $p(\phi \mid \beta)$ is a Dirichlet distribution and $p(z_s \mid \phi, D)$ is a multinomial distribution. Using similar techniques of Heinrich [21], we can get $p(z_s = k \mid D^{-s}) = \prod_{k=1}^{K} \frac{\triangle(\overrightarrow{n_k} + \beta)}{\triangle(\beta)}$, where $\overrightarrow{n_k} = \{n_k^w\}_{t=1}^V$, n_k^w denotes the number of times that word *w* has been observed in topic *k*. Here, we adopt the $\triangle$ function in Heinrich (2009), and we can have $\triangle(\beta) = \frac{\prod_{W=1}^{V}\Gamma(\beta)}{\Gamma(V\beta)}$ and $\triangle(\overrightarrow{n_k} + \beta) = \frac{\prod_{w \in d_s}\Gamma(n_k^w + \beta)}{\Gamma(n_k + V\beta)}$, where Γ denotes the gamma function.

Therefore, the full conditional distribution in Equation 2 for a short text can be derived as follows:

$$\begin{aligned} p(z_s = k \mid D^{-s}, \beta) &\propto \frac{p(z_s \mid D, \beta)}{p(z_s \mid D^{-s}, \beta)} \\ &\propto \frac{\triangle(\overrightarrow{n_d} + \beta)}{\triangle(\overrightarrow{n_d^{-s}} + \beta)} \\ &\propto \frac{\prod_{w \in s}\Gamma(n_k^w + \beta)}{\prod_{w \in s}\Gamma(n_k^{w,-s} + \beta)} \frac{\Gamma(n_k^{-s}) + V\beta}{\Gamma(n^k) + V\beta} \\ &\propto \frac{\prod_{w \in s}(n_k^{w,-s} + \beta)}{\prod_{i=1}^{N_s}(n_k^{-s} + V\beta)} \end{aligned} \quad (3)$$

For long texts, we will use the LDA model based on Gibbs Sampling [3]. The derivation of the conditional probability $p(z_{li} = k \mid D^{-li}, \alpha, \beta)$ is shown in paper [21].

HTTM requires tracking only small amounts of information from a corpus. For example, we need to keep track of a $K \times V$ (topics by words) matrix for ϕ, a $\mid L \mid \times K$ (long texts by topics) matrix for z_l, and an array of length $\mid S \mid$ (short texts) for z_s, where $\mid L \mid$ and $\mid S \mid$ are the number of long texts and short texts, respectively. We start this algorithm by assigning words of long texts to random topics and assigning short texts to random topics. Each iteration of the algorithm involves applying Equation 1 to every word in long texts, and applying Equation 2 to every text in short texts. The time complexity of each iteration in HTTM is O(K($\mid L \mid \overline{N_L}$ +$\mid S \mid \overline{N_S}$)), where $\overline{N_L}$ and $\overline{N_S}$ are the average length of long texts and short texts, respectively.

4. Experiments

In this section, we show the experimental results to demonstrate the effectiveness of our model by comparing it with three baselines on three datasets.

4.1. Datasets and Setup

Datasets: We choose the following two synthetic datasets and one real-word dataset to evaluate our model.

(1) Two synthetic datasets: *NIPS* [22] and *20 news group* [23]. The two datasets consist of long texts. For the purpose

TABLE 2. DATASET STATISTICS

(*S*: the number of short texts, *L*: the number of long texts, A_S: the average words of each short text, A_L: the average words of each long text, *V*: the size of words)

Dataset	S	L	A_S	A_L	V
NIPS	125454	174	10	786	15755
20News	33807	753	10	159	14420
News&Tweet	1137	614	7	264	15765

of evaluation, we randomly choose a certain percentage of documents as long texts, and generate a set of short texts by splitting the remaining documents into sentences, denoted as heterogeneous texts, which include short texts and long texts. Through splitting long texts into long texts and short texts, it is very useful for evaluation. We will explain it in the Evaluation Metrics section below.

(2) One real-word dataset: *News&Tweet*. We choose eight hot issues on January 05, 2016, which are "Tom Coughlin", "Oculus Rift", "SpaceX Rocket", "Donald Trump", "Windows 10", "Craig Strickland", "Bill Cosby", and "China stocks", respectively. We crawl news articles from different news websites, as long texts, and collect the tweets for each issue, as short texts.

For each dataset, we conduct the following preprocessing: (1) Convert all letters into lowercase; (2) Remove non-latin characters and stop words; (3) Remove words whose lengths are smaller than 3 or large than 20. In addition, after preprocessing we only choose these tweets whose length is larger than 5. Their statistics are summarized in Table 2.

Evaluation Metrics: A lot of metrics have been proposed for measuring the coherence of topics in texts [24], [25]. Although some metrics tend to be reasonable for long texts, they can be problematic for short texts [14]. Most conventional metrics try to estimate the likelihood of held-out testing data based on parameters inferred from training data. However, this likelihood is not necessarily a good indicator of the quality of extracted topics [26]. To provide a good evaluation, we evaluate all models from many aspects using different metrics,

(1) *Purity*: Purity is a new metric proposed by paper [14] which measures the coherence between the discoverable topics and the gold-standard topics. Since NIPS and 20 News are composed of long texts, LDA can achieve good results on the original datasets. Therefore, we view the topics extracted from the original NIPS and 20 News datasets with LDA as gold-standard. We can get the purity score through selecting a set of *T* top words from each topic respectively, and compare the set of words with those from gold-standard topics.

$$Purity = \frac{1}{TK}\sum_{i} max_j \mid \Gamma_{z_i} \cap \Gamma_{g_j} \mid$$

where z_i is a topic extracted from heterogeneous texts, g_j is a topic from gold-standard topics, and Γ_{z_i} and Γ_{g_j} are the sets of the top T words from topics z_i and g_j.

(2) *Qualitative and Quantitative Evaluation* [27], [28]: First, we discuss some exemplar topics learned by the four methods on the News&Tweet dataset. Each topic is visualized by the top ten words. Then, we evaluate our model in a quantitative manner based on the coherence measure (CM) to assess how coherent the learned topics are. For the News&Tweet dataset, we cannot use Purity metric, because the existing methods cannot get gold-standard topics on this dataset. For each topic, we choose the top 10 candidate words and ask human annotators to judge whether they are relevant to the corresponding topic. To do this, annotators need to judge whether a topic is interpretable or not. If not, the 10 words of the topic are labeled as irrelevant; otherwise these words are identified by annotators as relevant words for this topic. Coherence measure (CM) is defined as the ratio between the number of relevant words and the total number of candidate words. In our experiments, four graduate students participated in the labeling process.

Comparison Methods: We compare our model HTTM[1] with two classes of baselines: (1) Two state-of-the-art short text topic models, DMM [12] and BTM [11]. DMM uses the simple assumption that each text is sampled from only one latent topic. BTM learns topics by directly modeling the generation of word co-occurrence patterns in the corpus. (2) Long text topic model, LDA [3] which is the most widely used topic model.

For LDA, we use the package downloaded from `http://www.arbylon.net/projects/`. For BTM[2], we use the tools released by the authors. For DMM, we implement its code since the authors did not release the code.

Our model and the three comparing models are all based on Gibbs Sampling. Parameters of all these models are set as follows. First, the number of iterations is set to 2000, which is generally sufficient for convergence. Then, the two hyperparameters α and β are chosen according to their original paper. In LDA and BTM, α and β are set as 50/*K* and 0.01. In DMM and HTTM, α and β are set as 0.1 and 0.1. The percentage of long texts of the total documents in NIPS and 20 News is set to 0.1. The length threshold is 20, namely if the length of a text is larger than 20, it is treated as a long text, else as a short text.

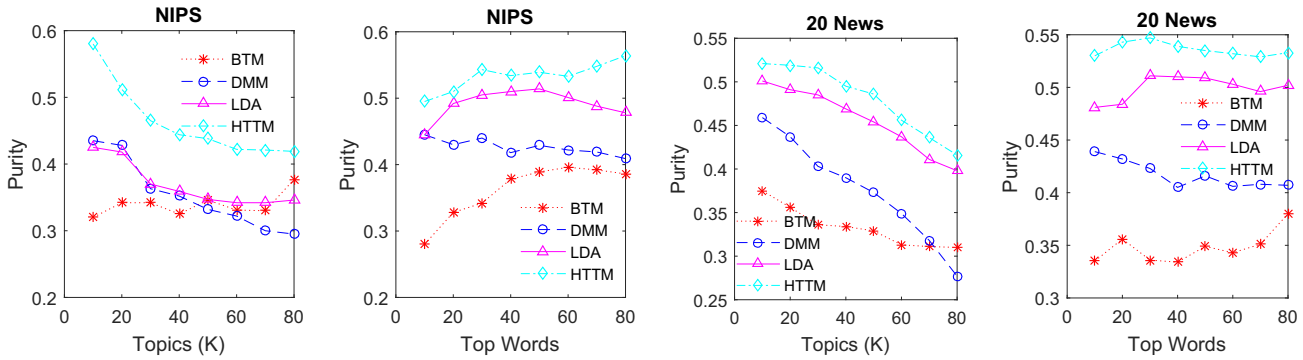
4.2. Purity

The number of latent topics *K* and top words could affect the performance of topic models. Therefore, for analyzing their impact on topic models, we carry out experiments by varying *K* and varying top words. Topic numbers *K* from 10 to 80 are investigated when *T*=20. Top words from 10 to 80 are investigated when *K*=20. The results are shown in Figure 2.

From Figure 2, we can see that the performance of HTTM is significantly better than other methods. These results suggest that any one of the two assumptions (one text contains one topic or multiple topics) does not work very

1. The source code can be downloaded at https://github.com/qiang2100/HTTM.git

2. https://github.com/xiaohuiyan/BTM



(a) Top 20 words (b) Topics=20 (c) Top 20 words (d) Topics=20

Figure 2. Experimental results on NIPS(left) and 20 news(right) data

well in heterogeneous texts. If we only use the assumption that one text contains only one topic, the models will have poor performance on long texts. On the other hand, if we only use the assumption that one text contains multiple topics, the models will not work well for short texts.

4.3. Qualitative Evaluation

Table 3 shows some exemplar topics learned by the four models on the News&Tweet dataset. Each topic is visualized by the top ten words. Words that are noisy and lack of representativeness are highlighted in bold. From Tabel 3, our model HTTM can learn more coherent topics with fewer noisy and meaningless words than all baselines. Long text topic model (LDA) that models each text as a mixture of topics is not fit for short texts, as a short text suffers from the sparsity of word co-occurrence patterns. Consequently, the top 10 words of "Oculus Rift" and "Windows 10" are not relevant to the corresponding topic. DMM has the worst results based on a simple assumption that each text is generated by one topic, because it is unreasonable to model a long text to contain only one topic. BTM posits that unordered word-pair co-occurring in a short text shares the same topic drawn from a mixture of topics cannot help improve the coherence of topic modeling since BTM ignores the knowledge that all the words in one topic have a high probability from the same topic. Simultaneously, we also show two additional results: LDA on only long texts and DMM on only short texts. Due to less information on each part of corpus, DMM and LDA have worst results than the results on the whole corpus. In conclusion, we can see that the topics learned by our model are far better than those learned by the baselines by integrating these two assumptions that each long or short text is either represented as the mixing proportions of multiple latent topics or one latent topic.

4.4. Quantitative Evaluation

Table 4 shows the coherence measure of topics inferred on the News&Tweet dataset. We can see our model HTTM significantly outperforms the baseline models. HTTM achieves an average coherence measure of 73%, which is larger than long text topic model (LDA) with a large margin. Compared to short text topic models (DMM and BTM), HTTM still has a big improvement. Our model produces better results compared to the baselines, which demonstrate the effectiveness of our model in heterogeneous texts HTTM which can also help improve the quality of topic modeling.

4.5. Clustering

We further compare the performance of all models in clustering. To provide alternative metrics, the normalized mutual information (NMI) is used to evaluate the quality of a clustering solution [12], [29]. NMI is an external clustering validation metric that effectively measures the amount of statistical information shared by random variables representing cluster assignments and user-labeled class assignments of the data points. NMI is always a number between 0 and 1, where 1 represents the best result and 0 means a random text partitioning. We run each model 20 times on each dataset and report the mean and standard deviation of their NMI values.

Table 5 shows the performance of all methods on the News&Tweet dataset. First, we can see that HTTM performs significantly better than LDA and DMM. It demonstrates that HTTM has the best performance for clustering in heterogeneous texts. This is because HTTM tackles heterogeneous texts by adopting different assumptions for short texts and long texts, and other topic models only use one assumption.

4.6. Influence of the Percentage of Long Texts

We also investigate the influence of the percentage of long texts on heterogeneous texts to the performance of all models when T=20 and K=20 for the NIPS dataset. Since more short texts (Twitter or Facebook) are generated online than long texts (News) in reality, the percentage of long texts in the whole texts is set from 0.05 to 0.6. From Figure

TABLE 3. TOPICS LEARNED FROM NEWS&TWEET DATASET

Topic	Method	Top Words
Tom Coughlin	LDA	coughlin giants coach season head **years** team york game super
	DMM	giants coughlin steps seasons **manage subscriptions alerts story artist charged**
	BTM	giants coughlin coach season head team **years** super game york
	HTTM	coughlin giants coach season head team **years** york game super
Oculus Rift	LDA	**windows oculus microsoft rift devices company headset january lumia hours**
	DMM	oculus rift **video** open headset reality free preorders **launch** virtual
	BTM	**windows microsoft oculus rift devices company news headset hours lumia**
	HTTM	oculus rift headset virtual reality january wednesday company touch price
SpaceX Rocket	LDA	rocket spacex falcon space landing launch musk stage **company** satellites
	DMM	**windows cosby year trump microsoft rocket china monday spacex time**
	BTM	rocket spacex falcon space launch landing musk stage **company** satellites
	HTTM	rocket spacex falcon space landing launch musk stage **company** satellites
Donald Trump	LDA	trump campaign donald iowa cruz republican clinton state hampshire states
	DMM	**coughlin bill cosby giants trump donald odell beckham coach video**
	BTM	trump campaign iowa cruz republican donald hampshire clinton rubio states
	HTTM	trump campaign donald iowa cruz republican clinton state hampshire states
Windows 10	LDA	**email comments post news today comment account facebook access badge**
	DMM	windows microsoft **keys** mobile **encryption** tablet users **acer liquid jade**
	BTM	**died year attack saudi january oregon fire iran star best**
	HTTM	windows microsoft devices lumia company mobile percent users running december
Craig Strickland	LDA	strickland craig body oklahoma singer missing morland country lake helen
	DMM	strickland craig singer missing country hope **great** dead church **friend**
	BTM	strickland craig body oklahoma morland missing singer lake helen country
	HTTM	strickland craig body oklahoma singer missing morland country lake monday
Bill Cosby	LDA	cosby bill women assault case camille sexual court constand charged
	DMM	**giants coughlin steps seasons manage subscriptions alerts story artist charged**
	BTM	cosby bill women case assault camille sexual court constand **years**
	HTTM	cosby bill women assault case camille sexual court **years** constand
China stocks	LDA	china stocks trading markets percent **year** market stock chinese investors
	DMM	china stocks hong kong policy data optimism open inflation **morning**
	BTM	china percent trading markets **year** stocks market stock chinese investors
	HTTM	china stocks trading markets percent **year** market stock chinese investors

TABLE 4. CM (%) ON NEWS&TWEET DATASET

Method	Annotator1	Annotator2	Annotator3	Annotator4	Mean	Standard Deviation
LDA	70	61	55	57	60.7	6.6
DMM	34	30	35	40	34.7	4.1
BTM	56	47	56	56	53.7	4.5
HTTM	**72**	**73**	**74**	**73**	**73**	**0.8**

TABLE 5. NMI VALUES ON NEWS&TWEET DATASET

Method	NMI
LDA	0.8350 ± 0.0349
DMM	0.5194± 0.0364
BTM	0.8786 ± 0.0281
HTTM	**0.9235 ± 0.0303**

3, we can see that HTTM outperforms other models. When increasing the value of the percentage, the purity of HTTM and LDA grows, and the purity of DMM and BTM declines. This is because the corpus includes more long texts when the percentage is large. But, when the percentage is small, LDA is worse than DMM. This shows that LDA is unsuitable for short texts.

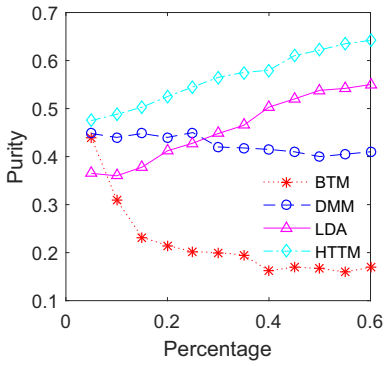


Figure 3. Comparison of the models with different percentages of long texts for NIPS dataset

5. Conclusion

We formulated a new problem of topic discovery from heterogeneous texts that include both short texts and long texts, which is challenging as existing topic models cannot work well on heterogeneous texts. We presented a collapsed Gibbs Sampling algorithm, Heterogeneous Text Topic Modeling (HTTM), by jointly considering the advantages of

short text topic modeling and long text topic modeling. The key idea is to adopt the simple assumption that each short text contains only one topic and the complex assumption that each long text contains multiple topics. Experimental results on real-world datasets demonstrated the substantial superiority of our HTTM model over the state-of-the-art methods.

Acknowledgement

This research is partially supported by the National 973 Program of China (No. 2013CB329604), the National Natural Science Foundation of China (No. 61229301, No. 61503116), and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China (No. IRT13059).

References

[1] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[3] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[4] T. Wang, V. Viswanath, and P. Chen, "Extended topic model for word dependency," in *ACL: short paper*, pp. 506–510, 2015.

[5] J.-P. Qiang, P. Chen, W. Ding, F. Xie, and X. Wu, "Multi-document summarization using closed patterns," *Knowledge-Based Systems*, vol. 99, pp. 28–38, 2016.

[6] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, pp. 80–88, ACM, 2010.

[7] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models:\# twitter trends detection topic model online.," in *COLING*, pp. 1519–1534, 2012.

[8] X. Wang, Y. Wang, W. Zuo, and G. Cai, "Exploring social context for topic identification in short and noisy texts," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[9] X. Yan, J. Guo, Y. Lan, J. Xu, and X. Cheng, "A probabilistic model for bursty topic discovery in microblogs.," in *AAAI*, pp. 353–359, 2015.

[10] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*, pp. 338–349, 2011.

[11] X. Cheng, X. Yan, Y. Lan, and J. Guo, "Btm: Topic modeling over short texts," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 12, pp. 2928–2941, 2014.

[12] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 233–242, 2014.

[13] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 889–892, 2013.

[14] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and sparse text topic modeling via self-aggregation," in *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 2270–2276, 2015.

[15] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270, 2010.

[16] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, ACM, 1999.

[18] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 306–315, 2004.

[19] J. Tang, M. Zhang, and Q. Mei, "One theme in all views: modeling consensus topics in multiple contexts," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 5–13, 2013.

[20] X. Wu, J.-P. Qiang, and F. Xie, "Pattern matching with flexible wildcards," *Journal of Computer Science and Technology*, vol. 29, no. 5, pp. 740–750, 2014.

[21] G. Heinrich, "Parameter estimation for text analysis," *University of Leipzig, Tech. Rep*, 2008.

[22] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," in *Advances in neural information processing systems*, pp. 497–504, 2004.

[23] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization.," tech. rep., DTIC Document, 1996.

[24] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108, 2010.

[25] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272, Association for Computational Linguistics, 2011.

[26] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, pp. 288–296, 2009.

[27] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling," *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2013.

[28] P. Xie, D. Yang, and E. P. Xing, "Incorporating word correlation knowledge into topic modeling," in *Conference of the North American Chapter of the Association for Computational Linguistics*, 2015.

[29] R. Huang, G. Yu, Z. Wang, J. Zhang, and L. Shi, "Dirichlet process mixture model for document clustering with feature partition," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 8, pp. 1748–1759, 2013.