Enhancing Topic Modeling on Short Texts with Crowdsourcing

Xiaoyan Yang XIAOYAN.YANG@ADSC.COM.SG
Advanced Digital Sciences Center

Shanshan Ying[*] YINGSHANSHAN@HUAWEI.COM
Huawei Technologies

Wenzhe Yu[†] YUWENZHE@QIYI.COM
iQIYI

Rong Zhang RZHANG@SEI.ECNU.EDU.CN
East China Normal University

Zhenjie Zhang ZHENJIE@ADSC.COM.SG
Advanced Digital Sciences Center

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

Topic modeling is nowadays widely used in text archive analytics, to find significant topics in news articles and important aspects of product comments available on the Internet. While statistical approaches, e.g. Latent Dirichlet Allocation (LDA) and its variants, are effective on building topic models on long texts, it remains difficult to identify meaningful topics over short texts, e.g. news titles and social-media messages. With the emergence and prosperity of crowdsourcing platforms, it becomes possible and easier for analytical systems to incorporate human intelligence into text analytics.

Different from traditional active learning techniques, the combination of crowdsourcing and machine learning poses new challenges on the design of simple tasks for non-experts to finish in seconds. In this paper, we design a new topic modeling technique, fully exploiting the basic intuitions of humans on short text reading. By requesting human labors to subjectively measure the similarity between short text pairs, the accuracy of the topic modeling algorithms could be greatly enhanced, regardless of the prior used in the graphical model. We present well-designed short text pair selection strategies for crowdsourcing and provide analysis on the convergence property of the inference algorithm. Empirical studies show that our proposed approaches improve the result topics on English tweets and Chinese microblogs, by requesting only a small number of labels from crowd.

Keywords: Topic Modeling, Crowdsourcing, Text Classification

1. Introduction

Topic modeling is nowadays widely used in text analysis with successful applications, such as finding trendy news topics from news articles and summarizing important aspects of restaurants from online comments. Statistical approaches, e.g. Latent Dirichlet Alloca-

[*] This work was done when the author was with Advanced Digital Sciences Center.
[†] This work was done when the author was an intern with Advanced Digital Sciences Center.

tion (LDA) and its variants (Blei et al., 2003, 2010), are proven effective on building topic models on archive of long texts, e.g. news articles and research paper abstracts. Given sufficiently long texts, these approaches are capable of identifying significant topics based on the co-occurrence relationship among words. However, it remains challenging to find meaningful topics over short texts, e.g. news titles, social-media messages and image captions. Such difficulties mainly come from the insufficient statistical significance on the word co-occurrence, when meaningful word combinations do not appear much more frequently than random word pairs. This limitation hinders the full utilization of such short text data, rendering poor topic results, e.g., on emerging social computing domain (Ramage et al., 2010).

Crowdsourcing is recently recognized as an effective tool to incorporate human intelligence into learning and analytics in an economically efficient way (Gomes et al., 2011; Tamuz et al., 2011; Yi et al., 2012). By decomposing a complicated problem into small Human Intelligence Tasks (HITs), the back-end learning and analytical algorithms are designed to utilize these answers from the crowd to these HITs, in order to improve the accuracy of learning outcomes. In traditional active learning, it is common to assume that the result labels come from domain experts, with desirable accuracy and reliability. Learning based on crowdsourcing, however, adopts completely different strategies, as workers on the crowdsourcing platform are only willing to pick simple HITs and their answers vary dramatically in terms of accuracy and response time.

To tackle the new challenges and fully exploit the power of human intelligence, this paper proposes a new general framework to enhance topic modeling with crowdsourcing. In existing studies on supervised or semi-supervised approaches for topic modeling, domain experts are expected to contribute topic labels on the texts (Blei and McAuliffe, 2007; Lacoste-Julien et al., 2008; Perotte et al., 2011; Ramage et al., 2009a, 2011) and correct meaningless word-topic associations (Andrzejewski et al., 2009; Hu et al., 2014). Such tasks are mostly inappropriate for crowdsourcing workers, with only basic text reading and understanding capabilities. In this paper, we present a new strategy which simply asks the crowd to give a binary label to indicate if the given two short texts are probably discussing the same topic. Such HITs are perfectly suitable for crowdsourcing workers, who could read the short texts and respond with a binary topic matching label within seconds. This approach is also fairly different from link-based labels used in text understanding, e.g. by retrieving the citation relationship among paper abstracts (Nallapati and Cohen, 2008), since citation edges sometimes may not reflect the topic similarity clearly.

Given the answers retrieved from the crowdsourcing platform, it is crucial to merge the noisy results into the existing statistical inference methods based on probabilistic graphical model. To accomplish seamless integration, we revise the original graphical model of LDA, including human intelligence as variables connected to the actual models of the texts, and deriving new Gibbs sampling algorithms in response to the model change. To maximize the economical efficiency of the topic modeling analysis, we also present an effective HIT selection strategy, which aims to minimize the variance on the latent variables in the inference algorithm. Our empirical studies on two real-world short text datasets, *Twitter* and *Weibo*, verify the significant improvement of our proposal over the standard collapsed Gibbs sampling method based on text data only. The major contributions of the paper are:

1. We present a new crowdsourcing framework for text topic modeling, which is sufficiently simple for workers and effective to enhance topic modeling accuracy.
2. We devise efficient inference algorithms to combine LDA-based text analytics and labels from crowdsourcing workers.
3. We provide rigorous theoretical analysis on the convergence property of the inference algorithm, to prove the efficiency of our proposal.
4. We propose a variance-based approach to select HITs to better utilize the human intelligence on crowdsourcing platform.
5. We evaluate our proposals with empirical studies, by comparing against state-of-the-art solutions in the literature, on two real world datasets.

2. Related Work

Knowledge about document corpora is available in various forms including categories and ratings. There has been a lot of work (Blei and McAuliffe, 2007; Lacoste-Julien et al., 2008; Ramage et al., 2009a, 2011; Perotte et al., 2011) on introducing knowledge to topic modeling. Modifications to LDA (Blei et al., 2003) are proposed to model such knowledge and document content together. Supervised LDA (Blei and McAuliffe, 2007) and DiscLDA (Lacoste-Julien et al., 2008) first model corpora with single label attached to each document. Labeled LDA (Ramage et al., 2009a) and PLDA (Ramage et al., 2011) are then proposed to model documents with multiple labels. Labels with a hierarchical structure are modeled in HSLDA (Perotte et al., 2011). Other topic modeling work like MM-LDA (Ramage et al., 2009b) also models documents and tags together. However, different from the above work, MM-LDA does not model the relationship between words and tags directly. Due to its underlying exchangeability assumption, MM-LDA allows topics that generate words to be different from those that generate tags. Rodrigues et al. (2015) extends supervised LDA (Blei and McAuliffe, 2007) by obtaining labels from the crowd. It models the reliability of multiple annotators so that it is able to capture both the true labels and the noisy labels from the crowd. However, such HITs required in Rodrigues et al. (2015) are more suitable for long documents with well-defined labels.

Other domain knowledge about associations of words with topics are put in topic models in (Andrzejewski et al., 2009), which introduces constraints on the composition of words in topics as “must-link” or “cannot-link”. For example, a must-link constraint requires a group of words to have similar probability within any topic. (Wang et al., 2016) explores user-specified keywords to guide aspect related topics discovery for focused analysis.

Obtaining knowledge of document corpora introduced in the above models usually require substantial effort from domain experts, and are thus not suitable for crowdsourcing workers. Our proposed HIT tasks, on the contrary, are very simple and require only basic reading and understanding capabilities. The binary label indicates whether two short documents are discussing the same topic. With sufficient crowdsourcing labels, the ratio between positive labels and total labels are expected to reflect the similarity of two documents.

Another line of related research is topic modeling work (Cohn and Hofmann, 2001; Erosheva et al., 2004; Nallapati and Cohen, 2008; Gruber et al., 2008; Mei et al., 2008; Nallapati et al., 2008) that models document corpora with network structure, as the crowdsourcing labels could be viewed as a special type of link information between documents. Such network structure reflects knowledge about relationship between documents. The problem of these models is that although they model words and links together, they fail to model the relationship between them explicitly, as they allow topics that generate words to be different from topics that generates links due to their underlying exchangeability assumptions. Another problem of (Cohn and Hofmann, 2001; Erosheva et al., 2004; Nallapati and Cohen, 2008; Gruber et al., 2008) is that they treat links similarly to words but from a different vocabulary, and thus cannot be generalized to new documents outside of the training data (Chang and Blei, 2009). HTM (Sun et al., 2008) also introduces network structure into topic modeling by allowing words to be generated by topics of cited documents. But it has the same problem as above in generalizing to new documents.

The relationship between document content and links are explicitly modeled in (Chang and Blei, 2009; Chen et al., 2015; Du et al., 2015). Chang and Blei (2009) propose Relational Topic Model (RTM), in which a link between two documents is generated based on their topic assignments to words. Chen et al. (2015) extends RTM to asymmetric network. Our crowdsourcing LDA (cLDA) model is most related to RTM, in which symmetric relation of documents is assumed. But different from RTM, cLDA models links (crowdsourcing labels) to be related to topic distributions of documents. Topic-Link LDA (Liu et al., 2009) further introduces author community and models links to be related to topic distributions of documents and communities of authors. Du et al. (2015) combines LDA with constraints derived from relative similarities among documents, which act as a regularizer.

cLDA is also related to interactive topic modeling (ITM) (Hu et al., 2014) in continuously incorporating knowledge during inference. But the type of knowledge accepted by ITM is different from that of cLDA. ITM can interactively add constraints as in (Andrzejewski et al., 2009) by doing selective state unassignment during inference whereas cLDA can allow crowdsourcing labels to be added iteratively during inference.

3. Preliminaries

Following the standard model of LDA (Blei et al., 2003), the generation of the text documents follows the simulated process described below. Given a specific number of topics k, a global dictionary $\boldsymbol{\phi}$ is created, such that each topic is associated with a mutli-nominal distribution $\boldsymbol{\phi}_j$ on the words with prior β. For each document d_i, a topic distribution $\boldsymbol{\theta}_i$ is generated by drawing a sample from Dirichlet distribution with prior α. When writing a particular word l in the document d_i, the process first chooses a topic z_{il} based on the distribution $\boldsymbol{\theta}_i$. It then picks up a word using the probability vector $\boldsymbol{\phi}_{z_{il}}$ in the global dictionary $\boldsymbol{\phi}$. This generation model is almost equivalent to probabilistic Latent Semantic Index model (Hoffman, 1999), except on the Dirichlet prior on the topic distributions.

Based on the general text generation model described above, the goal of topic modeling analysis is to recover the dictionaries, i.e. the multi-nominal distributions $\{\boldsymbol{\phi}_j\}$ for all topics, by running the inference algorithm over the text sequences, $\{d_1, d_2, \ldots, d_N\}$. When there is no additional information available on the topics, inference approaches, e.g. Gibbs sampling and variational inference, are applied on the text to build the dictionaries, under the objective of posterior likelihood maximization.

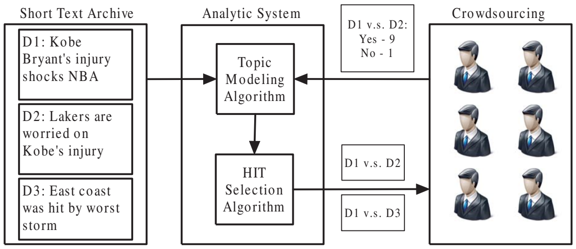

Short Text Archive
D1: Kobe Bryant's injury shocks NBA
D2: Lakers are worried on Kobe's injury
D3: East coast was hit by worst storm
Analytic System
Topic Modeling Algorithm
HIT Selection Algorithm
D1 v.s. D2: Yes - 9 No - 1
D1 v.s. D2
D1 v.s. D3
Crowdsourcing

Figure 1: The general architecture of the system

3.1. System Framework

In Figure 1, we present the general architecture of our crowdsourcing-based topic modeling framework. The short text archive on the left side consists of N short text documents, denoted by $\{d_1, d_2, \ldots, d_N\}$. These texts are fed into the topic modeling algorithm for analysis. Based on the analysis results, the HIT selection algorithm generates a group of HITs, each of which contains a pair of short texts, e.g. (d_i, d_r). These HITs are submitted to the crowdsourcing platform, on which workers could download the short text pairs and attach binary labels to these pairs. The worker is supposed to answer 0 if he believes these two short texts are dissimilar on the underlying topic, or answer 1 if the topics of the texts are sufficiently similar based on his understanding. In our example in Figure 1, the pair of (d_1, d_2) is presented to 10 independently chosen workers on the platform, with 9 positive answers and 1 negative answer. Obviously, the aggregation of these answers indicates these two short texts are similar on discussion topic, because both of the texts are about the injury of NBA basketball star Kobe Bryant. Such aggregation also helps reduce the uncertainty in labels provided by single crowdsourcing worker.

All of these results to the HITs are collected and aggregated by the crowdsourcing platform, such that the analysis algorithm knows the number of positive and negative answers over the selected short text pairs. The topic modeling algorithm recomputes the topics using both the original text archive and the binary answers from the crowd. When an updated model with new dictionaries is generated, a new group of HITs are constructed and transmitted to the crowdsourcing platform for additional human intelligence. A few round of iterations could be run until the topic modeling algorithm is confident on the convergence of the modeling output.

3.2. Algorithm Framework

The core of the system is the inference algorithm, which generates the topic dictionaries, based on the short texts and the answers to the HITs from crowd. It is not trivial to extend the existing inference algorithms for standard LDA, because it is not straightforward to directly incorporate the result labels on topic similarity into the generative model of LDA. To address this challenge, we revise the original graphical model of the generation mechanism, to formalize the intuition on topic similarity of crowdsourcing workers by the topics on the texts. Specifically, different from standard LDA model, we add new variables into the graphical model to represent human's understandings to the similarities between pairs of text documents. Let p_{ir} denote the number of positive feedbacks on document pair d_i and d_r. Similarly, n_{ir} denotes the number of negative answers on d_i and d_r retrieved from the

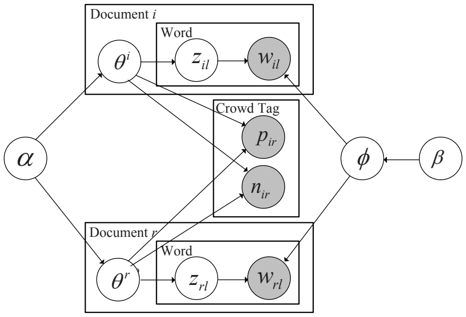


Figure 2: The graphical model on the variables underlying our generative model.

crowdsourcing platform. To affect the topic distribution in d_i and d_r, we divide each p_{ir} into k parts, i.e. $p_{ij,rj}$ for each topic j. Each $p_{ij,rj}$ indicates the number of crowdsourcing workers agreeing that both d_i and d_r discuss on topic j. Different from p_{ir}, we divided n_{ir} into $k^2 - k$ parts, such that each $n_{ij,rh}$ is an estimated number of answers with d_i and d_r associated with topic j and h respectively. In Figure 2, we present a complete probabilistic graphical model of the generation process, in which observable variables are marked in grey.

Mathematically, the generation process could be formulated as drawing samples following the probability distribution functions below,

$$\Pr(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{p}, \boldsymbol{n}, \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \Pr(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta}) \Pr(\boldsymbol{z}|\boldsymbol{\theta}) \Pr(\boldsymbol{p}, \boldsymbol{n}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}|\boldsymbol{\alpha}) \tag{1}$$

in which $\Pr(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\beta})$, $\Pr(\boldsymbol{z}|\boldsymbol{\theta})$ and $\Pr(\boldsymbol{\theta}|\boldsymbol{\alpha})$ follow the distributions in the original LDA model, while $\Pr(\boldsymbol{p}, \boldsymbol{n}|\boldsymbol{\theta})$ follows the distribution,

$$\Pr(\boldsymbol{p}, \boldsymbol{n}|\boldsymbol{\theta}) = \prod_i \prod_{r \neq i} \Pr(p_{ir}, n_{ir}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_r) \tag{2}$$

Intuitively, $\Pr(p_{ir}, n_{ir}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_r)$ indicates the distribution of positive and negative crowd labels on a pair of short texts d_i and d_r. Based on the assumption on the consistency between human's understanding and topic distribution, we further build a mathematical connection from $(\boldsymbol{\theta}_i, \boldsymbol{\theta}_r)$ to $\Pr(p_{ir}, n_{ir}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_r)$. Given k topics in the archive, with *independent* topic distribution vectors $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ik})$ and $\boldsymbol{\theta}_r = (\theta_{r1}, \theta_{r2}, \ldots, \theta_{rk})$, we define a coherence matrix to model when people happen to sense the texts under a particular pair of topics,

$$\boldsymbol{\theta}_i^T \boldsymbol{\theta}_r = \begin{pmatrix} \theta_{i1}\theta_{r1} & \theta_{i1}\theta_{r2} & \ldots & \theta_{i1}\theta_{rk} \\ \theta_{i2}\theta_{r1} & \theta_{i2}\theta_{r2} & \ldots & \theta_{i2}\theta_{rk} \\ \ldots & \ldots & \ldots & \ldots \\ \theta_{ik}\theta_{r1} & \theta_{ik}\theta_{r2} & \ldots & \theta_{ik}\theta_{rk} \end{pmatrix}$$

Given p_{ir}, n_{ir} and $\boldsymbol{\theta}_i^T \boldsymbol{\theta}_r$, we decompose the crowd labels based on the topic j short texts d_i and d_r probably share, as $p_{ij,rj} = \frac{\theta_{ij}\theta_{rj}}{\sum_j \theta_{ij}\theta_{rj}} p_{ir}$ and $n_{ij,rh} = \frac{\theta_{ij}\theta_{rh}}{\sum_{j \neq h} \theta_{ij}\theta_{rh}} n_{ir}$. The probability of observing crowd labels $\{p_{ir}, n_{ir}\}$ is thus finalized as

$$\Pr(p_{ir}, n_{ir}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_r) = \binom{p_{ir} + n_{ir}}{\{p_{ij,rj}\}, \{n_{ij,rh}\}} \prod_j \Big(\theta_{ij}\theta_{rj}\Big)^{p_{ij,rj}} \prod_j \prod_{h \neq j} \Big(\theta_{ij}\theta_{rh}\Big)^{n_{ij,rh}}$$

where the combination number is calculated by

$$\binom{p_{ir}+n_{ir}}{\{p_{ij,rj}\},\{n_{ij,rh}\}} = \frac{(p_{ir}+n_{ir})!}{\prod_j p_{ij,rj}! \prod_j \prod_{h\neq j} n_{ij,rh}!}.$$

Notice that $p_{ij,rj}$ and $n_{ij,rh}$ are rounded up to ensure $\Pr(p_{ir}, n_{ir}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_r)$ follows a valid multinomial distribution in practice, which is important to the validity of inference algorithm in the next section.

4. Inference Approach

Based on the algorithm framework described in previous section, we design a new inference algorithm to handle the additional variables in the graphical model and HIT answers from crowdsourcing. In this section, we simply assume the selection of HITs are done by black box. We will discuss HIT selection strategy in later sections to fill the gap.

We revise the collapsed Gibbs sampling method (Griffiths and Steyvers, 2004) to infer the latent variables in the graphical model. The sampling equations for our new model is different from that for standard LDA, mainly on the new labels from the crowd, i.e. $\boldsymbol{p}$ and $\boldsymbol{n}$. It is thus difficult to collapse the impact of $\boldsymbol{\theta}$ during Gibbs sampling. We solve the problem by dividing each Gibbs sampling iteration into two steps, sampling on the probability on z_{il} and sampling on the probability of $\boldsymbol{\theta}_i$. The additional sampling step brings more complexity to the inference procedure, but simplifies the processing logic by isolating the impacts of $\boldsymbol{p}$ and $\boldsymbol{n}$ from the other part of the graphical model in Figure 2.

4.1. Sampling on Word-Topic Association

The first sampling step focuses on the word-topic association, i.e. the topic z_{il} on each word w_{il}. Given the observations on $\boldsymbol{w}$, $\boldsymbol{p}$, $\boldsymbol{n}$, $\boldsymbol{\theta}$ and topics of other words $\boldsymbol{z_{-il}}$, the target of the sampling in this step is to estimate the posterior probability of $z_{il} = j$ conditioned on the observations. Therefore, the sampling probability for a topic for the lth word w_{il} in document d_i in our new model is derived as

$$\Pr(z_{il}=j|\boldsymbol{z}_{-il},\boldsymbol{w},\boldsymbol{p},\boldsymbol{n},\boldsymbol{\theta}) \propto \Pr(w_{il}|z_{il}=j,\boldsymbol{z}_{-il},\boldsymbol{w}_{-il},\boldsymbol{p},\boldsymbol{n},\boldsymbol{\theta}) \Pr(z_{il}=j|\boldsymbol{z}_{-il},\boldsymbol{w}_{-il},\boldsymbol{p},\boldsymbol{n},\boldsymbol{\theta}) \quad (3)$$

in which $\Pr(w_{il}|z_{il} = j, \boldsymbol{z}_{-il}, \boldsymbol{w}_{-il}, \boldsymbol{p}, \boldsymbol{n}, \boldsymbol{\theta}) = \Pr(w_{il}|z_{il} = j, \boldsymbol{z}_{-il}, \boldsymbol{w}_{-il}) = \frac{N^{(w_l)}_{-l,j}+\beta}{N^{(\cdot)}_{-l,j}+W\beta}$[1]and $\Pr(z_{il} = j|\boldsymbol{z}_{-il}, \boldsymbol{w}_{-il}, \boldsymbol{p}, \boldsymbol{n}, \boldsymbol{\theta}) = \Pr(z_{il} = j|\theta_i) = \theta_{ij}$.

The computational complexity of this sampling step is linear to the number of words. Therefore, the sampling on word-topic association is updated in a very efficient way.

4.2. Sampling on Document-Topic Association

In the second sampling step, we fix word-topic associations $\{z_{il}\}$ as observable variables and turn to sample on document-topic associations $\{\theta_{ij}\}$. The sampling probability for topic distribution $\boldsymbol{\theta}_i$ of document d_i is derived as

$$\Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i},\boldsymbol{z},\boldsymbol{w},\boldsymbol{p},\boldsymbol{n}) = \Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i},\boldsymbol{z},\boldsymbol{p},\boldsymbol{n}) \propto \Pr(\boldsymbol{z}_i|\boldsymbol{\theta}_i,\boldsymbol{\theta}_{-i},\boldsymbol{z}_{-i},\boldsymbol{p},\boldsymbol{n}) \Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i},\boldsymbol{z}_{-i},\boldsymbol{p},\boldsymbol{n}) \quad (4)$$

[1] Notice that ϕ is collapsed and is thus omitted.

In the formula above, we have $\Pr(\boldsymbol{z}_i|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{p}, \boldsymbol{n}) = \Pr(\boldsymbol{z}_i|\boldsymbol{\theta}_i) = \prod_l p(z_{il}|\boldsymbol{\theta_i}) = \prod_j \theta_{ij}^{m_{ij}}$, where m_{ij} is the number of words in document d_i that has been assigned to topic j, and

$$\begin{aligned}\Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{p}, \boldsymbol{n}) = \Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{p}, \boldsymbol{n}) \quad &\propto \quad \Pr(\boldsymbol{p}_i, \boldsymbol{n}_i|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}, \boldsymbol{p}_{-i}, \boldsymbol{n}_{-i}) \Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{p}_{-i}, \boldsymbol{n}_{-i}) \\ &= \quad \Pr(\boldsymbol{p}_i, \boldsymbol{n}_i|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}) \Pr(\boldsymbol{\theta}_i)\end{aligned}$$

In particular, the probability $\Pr(\boldsymbol{p}_i, \boldsymbol{n}_i|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i})$ is estimated as $\prod_r \Pr(\boldsymbol{p}_{ir}, \boldsymbol{n}_{ir}|\boldsymbol{\theta}_i, \boldsymbol{\theta}_r) = C_i \prod_j \theta_{ij}^{h_{ij}}$, $C_{ir} = \binom{p_{ir}+n_{ir}}{\{p_{ij,rj}\},\{n_{ij,rh}\}}$, with

$$C_i = \prod_r C_{ir} \prod_j \prod_r (\theta_{rj}^{\sum_r (p_{ij,rj})} \prod_h \theta_{rh}^{\sum_h n_{ij,rh}})$$

and $h_{ij} = \sum_r (p_{ij,rj} + \sum_h n_{ij,rh})$.

Since $\Pr(\theta_i)$ is Dirichlet(α) and conjugate to stochastic process $\Pr(\boldsymbol{p}_i, \boldsymbol{n}_i|\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i})$, the posterior probability follows the distribution $\Pr(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}, \boldsymbol{z}_{-i}, \boldsymbol{p}, \boldsymbol{n})$ will be Dirichlet($\alpha + h_{i1}, \ldots, \alpha + h_{ik}$). By further exploring the conjugacy in Equation (4), we finally reach the sampling rule:

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\alpha + h_{i1} + m_{i1}, \ldots, \alpha + h_{ik} + m_{ik}) \tag{5}$$

The computational complexity of this sampling step is $O(CNLk^2)$, where N is the total number of documents, L is the maximum number of times a document can be selected in HITs, and C is a constant. Since $L < k^2 \ll N$, the computational complexity is linear to N.

4.3. Convergence Analysis

Based on the sampling distribution of $\boldsymbol{\theta}$ used in Equation (5), the inference result $\boldsymbol{\theta}_i$ converges to its expectation, when there are more crowd labels related to d_i available. In particular, due to the property of Dirichlet process, the expectation of θ_{ij} is

$$\boldsymbol{E}(\theta_{ij}) = \frac{\alpha + h_{ij} + m_{ij}}{k\alpha + \sum_l h_{il} + \sum_l m_{il}}.$$

This section devotes to understanding the convergence rate of the inference algorithm, in terms of the amount of crowd efforts.

Assume θ_{ij}^* is the actual probability of d_i in topic j as well as the extreme value of θ_{ij}, i.e. $\theta_{ij}^* = \lim_{\lambda_i \to \infty} \theta_{ij}$. Firstly, the following lemma implies the direct connection between h_{ij} and θ_{ij}^*.

Lemma 1 *If λ_i is the number of retrieved crowd labels related to document d_i, we have $\lim_{\lambda_i \to \infty} \frac{h_{ij}}{\sum_l h_{il}} = \theta_{ij}^*$.*

Proof Since α is a constant prior specified by the user and $0 \leq \sum_l m_{il} \leq c_i$, when h_{ij} is sufficiently large, $\boldsymbol{E}(\theta_{ij}) = \frac{\alpha+h_{ij}+m_{ij}}{k\alpha+\sum_l h_{il}+\sum_l m_{il}} \to \frac{h_{ij}}{\sum_l h_{il}}$. Therefore, $\frac{h_{ij}}{\sum_l h_{il}}$ converges to θ_{ij}^*. ■

In other words, the crowd labels dominate the topic probability estimation when there are sufficient crowd labels related to d_i available to the analysis algorithm.

Lemma 2 *For any pair of* (d_i, d_r) *and any topic* j*, we always have*

$$\boldsymbol{E}\left(p_{ij,rj} + \sum_{h\neq j} n_{ij,rh}\right) = \theta^*_{ij} \sum_l \boldsymbol{E}\left(p_{il,rl} + \sum_{h\neq l} n_{il,rh}\right).$$

Proof To prove the lemma, we focus on the analysis on $\boldsymbol{E}\left(p_{ij,rj} + \sum_{h\neq j} n_{ij,rh}\right)$. Based on the matrix representation of independent topic distributions of d_i and d_r, the expectation of $p_i r$ and $n_i r$ are $\boldsymbol{E}(p_{ir}) = \left(\sum_j \theta_{ij}\theta_{rj}\right)\lambda_{ir}$, and $\boldsymbol{E}(n_{ir}) = \left(\sum_{j\neq h} \theta_{ij}\theta_{rh}\right)\lambda_{ir}$ respectively.

Combined the equations on $\boldsymbol{E}(p_{ir})$ and $\boldsymbol{E}(n_{ir})$ above, the expectation of $p_{ij,rj}$ and $n_{ij,rh}$ are $\boldsymbol{E}(p_{ij,rj}) = \theta_{ij}\theta_{rj}\lambda_{ir}$ and $\boldsymbol{E}(n_{ij,rh}) = \theta_{ij}\theta_{rh}\lambda_{ir}$. By inserting these expectations into the target equation of the lemma, the equality relationship between both hands is straightforward. ■

The most important implication of last lemma is that the estimated ratio of topic j on text d_i does not depend on the d_r paired in the human intelligence tasks. Therefore, the convergence of θ_{ij} does not rely on the assignment strategy on $\{\lambda_{i1}, \ldots, \lambda_{iN}\}$, which is used in the proofs of next theorem.

Theorem 3 *When the number of labels from crowd related to document* d_i *is* $\lambda_i \geq \frac{(2\theta^*_{ij}k+2)\alpha+(2\theta^*_{ij}+2)c_i}{\epsilon\theta^*_{ij}}$*, the error on the* θ_{ij} *is bounded with high probability,* $\Pr(|\theta^*_{ij} - \boldsymbol{E}(\theta_{ij})| \leq \epsilon) \geq 1 - 2\exp\left(-\frac{\epsilon^2}{2}\lambda_i\right)$.

Proof To prove the theorem, we introduce a new variable $\theta'_{ij} = \frac{h_{ij}}{\sum_l h_{il}}$. To prove the theorem, we prove the following two inequalities $\Pr(|\theta^*_{ij} - \boldsymbol{E}(\theta'_{ij})| \leq \epsilon/2)$ and $\Pr(|\boldsymbol{E}(\theta'_{ij}) - \boldsymbol{E}(\theta_{ij})| \leq \epsilon/2)$ separately.

The first probability is bounded by utilizing Hoeffding inequality. By Lemma 2, the contribution of each feedback from the crowd follows the same distribution, $\theta'_{ij} = \frac{h_{ij}}{\sum_l h_{il}}$ is thus equivalent to the sampling procedure from a binary distribution with expected probability $\Pr(X = 1) = \theta^*_{ij}$. By Hoeffding inequality, the probability of deviation larger than $\epsilon/2$ is no larger than $\Pr\left(\left|\theta^*_{ij} - \boldsymbol{E}(\theta'_{ij})\right| \geq \frac{\epsilon}{2}\right) \leq 2\exp\left(-\frac{\epsilon^2}{2}\lambda_i\right)$.

The second probability is bounded by eliminating the impact of small constants in the expectation of θ. Since $\frac{h_{ij}}{\sum_l h_{il}}$ converges to θ^*_{ij}, we aim to find condition on λ_i, such that the following inequality always holds: $\left(1 - \frac{\epsilon}{2}\right)\theta^*_{ij} \leq \frac{\alpha+\theta^*_{ij}\lambda_i+m_{ij}}{k\alpha+\lambda_i+\sum_l m_{il}} \leq \left(1 + \frac{\epsilon}{2}\right)\theta^*_{ij}$. For the left part of the inequality, by the fact $0 \leq m_{ij} \leq 1$ and $1 \leq \sum_l m_{il} \leq c_i$, it is valid when $\left(1 - \frac{\epsilon}{2}\right)\theta^*_{ij} \leq \frac{\alpha+\theta^*_{ij}\lambda_i}{k\alpha+\lambda_i+c_i}$. It leads to the condition

$$\lambda_i \geq \frac{\left(1 - \frac{\epsilon}{2}\right)\theta^*_{ij}\left(k\alpha + c_i\right) - \alpha}{\frac{\epsilon}{2}\theta^*_{ij}} \tag{6}$$

Similarly, on the right side of the inequality, we have $\frac{\alpha+\theta^*_{ij}\lambda_i+c_i}{k\alpha+\lambda_i+c_i} \leq \left(1 + \frac{\epsilon}{2}\right)\theta^*_{ij}$. which leads to another condition

$$\lambda_i \geq \frac{\alpha + c_i - \left(1 + \frac{\epsilon}{2}\right)\theta^*_{ij}\left(k\alpha + c_i\right)}{\frac{\epsilon}{2}\theta^*_{ij}} \tag{7}$$

Both of the conditions in Equation (6) and Equation (7) are satisfied when $\lambda_i \geq \frac{(2\theta^*_{ij}k+2)\alpha+(2\theta^*_{ij}+2)c_i}{\epsilon\theta^*_{ij}}$. Because $\theta^*_{ij} \leq 1$, the absolute error on $|\boldsymbol{E}(\theta'_{ij})-\boldsymbol{E}(\theta_{ij})|$ is no larger than $\frac{\epsilon}{2}$. Therefore, when the condition is met, the probability is $\Pr\left(|\boldsymbol{E}(\theta'_{ij}) - \boldsymbol{E}(\theta_{ij})| \leq \frac{\epsilon}{2}\right) = 1$. This completes the proof of the theorem. ■

Basically, the last theorem implies that the error converges to zero exponentially, when the number of crowd labels on a particular text d_i reaches a minimum threshold. When $\epsilon = 0.1$ and d_i is attached to a dominant topic $\theta^*_{ij} = 0.5$, for example, a few dozens of crowd labels are sufficient to control the error of the result topic on d_i within the error bound. However, note that the bound derived in previous lemma remains loose. The actual convergence rate of the estimation, as is shown in the empirical evaluation, is much sharper than the estimation.

5. HIT Selection Strategies

While the analysis in previous section has implied the exponential convergence of our analysis algorithm with extra knowledge from crowd labels, different HIT selection strategies may affect the performance of the topic modeling system in a variety of ways. In this section, we discuss two general HIT selection strategies – *uniform* and *varianced-based* approaches. **Uniform Selection** A straightforward solution to the selection of human intelligence tasks (HITs) is uniform sampling over all (d_i, d_r) pairs. This strategy is very simple to implement and helpful to inference algorithm in identifying the boundary between topics. However, uniform sampling may generate a large number of HITs for text pairs from different topics and almost no HIT for text pairs from the same topic. Since the local dictionaries for the topics are usually very sparse, these HITs may not be productive to refine local dictionaries by a significant margin. This phenomenon is formalized by the following lemma.

Lemma 4 *If the underlying Dirichlet process is run with hyperparameter α and topic number k, the expected probability of choosing a pair of text (d_i, d_r) from the same topic j by uniform sampling is $\frac{k-1}{k(k\alpha+1)} + \frac{1}{k}$.*

Proof Assume $(X_1, X_2, \ldots, X_k)$ are the variables generated by the Dirichlet process with parameter α. Topic j is chosen by the probability X_j. Therefore, if we randomly choose two texts, the probability of having them from the same topic is $\sum_j (X_j)^2$. The expected probability is thus $\mathbf{E}\left(\sum_j (X_j)^2\right) = \sum_j \mathbf{E}\,(X_j)^2$, by the linearity property of expectation.

Based on the property of Dirichlet process, the expectation of X_j is $\mathbf{E}(X_j) = \frac{1}{k}$ and the variance of X_j is $\mathbf{V}(X_j) = \frac{k-1}{k^2(k\alpha+1}$. Because $\mathbf{V}(X_j) = \mathbf{E}\,(X_j)^2 - (\mathbf{E}(X_j))^2$, for any topic j, we have

$$\mathbf{E}\,(X_j)^2 = \frac{k-1}{k^2(k\alpha+1)} + \frac{1}{k^2}.$$

By summing up over all topic j, we reach the conclusion of the lemma that the probability is exactly $\frac{k-1}{k(k\alpha+1)} + \frac{1}{k}$. ■

Based on the lemma above, the probability of generating HITs with text pairs from the same topic tends to $\frac{1}{k}$, when α increases. On large archives such as Twitter dataset, the number of topics k is usually not a small number. This further makes the problem worse. In the following subsection, we present a new solution by choosing text pairs based on the potential contribution to the model update.

Variance-Based Selection In this part of the section, we present a new guideline on HIT selection to minimize the variance on the latent variables during Gibbs sampling, such that the answers of these HITs could possibly minimize the uncertainty after another round of inference. Basically, our variance-based selection (VBS) guideline is motivated by the observation on Equation (3) such that

$$\boldsymbol{V}(z_{il}) \propto \sum_{j=1}^{k} \left(\frac{N_{-l,j}^{(w_l)} + \beta}{N_{-l,j}^{(\cdot)} + W\beta} \right)^2 \boldsymbol{V}\left(\theta_{ij}\right).$$

where $\boldsymbol{V}(\theta_{ij}) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0+1)}, \alpha_0 = \sum_j \alpha_j, \forall j \ \alpha_j = \alpha$. When document d_i is selected with some d_r in one HIT with crowd labels (p_{ir}, n_{ir}), according to Equation (5), we have $\alpha_j = \alpha + h_{ij} + m_{ij}$ and $\alpha_0 = \sum_j \alpha_j = k\alpha + \sum_j (h_{ij} + m_{ij}) = k\alpha + p_{ir} + n_{ir} + \sum_j m_{ij}$. Since $0 \leq h_{ij} \leq p_{ir} + n_{ir}$, crowd labels will help reduce the variance $\boldsymbol{V}(\theta_{ij})$ of θ_{ij}, which also help reduce $\boldsymbol{V}(z_{il})$. We therefore propose a variance-based HIT selection method that samples documents proportional to $\frac{\boldsymbol{V}(\theta_i)}{\sum_{i'} \boldsymbol{V}(\theta_{i'})}$.

6. Experiments

Datasets We test on two real datasets *Twitter* and *Weibo* to verify the effectiveness of our proposals. The *Twitter* dataset consists of tweets collected on Twitter posted on Jan 23, 2011. By filtering those without any stop word in English, we retrieve English-based tweets, which probably contain a small number of Spanish words. The *Weibo* dataset consists of microblogs mainly in Chinese, collected on Sina Weibo posted on Nov 10, 2014.

In preprocessing, we filter duplicate tweets/microblogs, and remove stop words and infrequently occurring words from the datasets. In particular, words appeared less than 5 times in the *Twitter* dataset and 10 times in the *Weibo* dataset are removed. English words are stemmed. After preprocessing, the *Twitter* dataset has 132,985 documents with 741,080 total number of words and 11,187 unique words. The *Weibo* dataset has 138,455 documents with 4,304,291 total number of words and 40,317 unique words.

Baseline Approach To the best of our knowledge, there is no existing topic modeling work that can handle crowdsourcing labels directly. As the label information can be viewed as a special type of link information indicating topic similarity between document pairs, we compare our model, denoted as cLDA, with RTM (Chang and Blei, 2009), which also models connections between documents and outperforms several baseline approaches in link prediction as shown in Chang and Blei (2009). In the experiments, we use the RTM model using collapsed Gibbs sampling released in (Chang, 2012), which implements the exponential link probability function and the EM algorithm that iteratively estimates the regression parameter.

To validate the effectiveness of our variance-based HIT selection strategy, we evaluate cLDA with different HIT selection strategies. Let cLDA-VAR denote the cLDA model using variance-based HIT selection approach. We also implement cLDA-UNI, which selects document pairs to label using uniform sampling, and cLDA-TAG, which selects document pairs containing same tags uniformly.

Crowdsourcing Labels We recruit graduate students as our crowdsourcing workers. To save manpower, a student is given a selected document pair (d_i, d_r) and is asked to distribute m tokens between two side, e.g. p_{ir} on *positive* and n_{ir} on *negative*, based on his/her understandings to short texts. Such a task simulates m HITs executed in real crowdsourcing platform, where m individual users are asked to answer Yes/No on whether two documents are similar in topics. In the experiments, m is set to 5. Each document pair (d_i, d_r) is assigned to three students and the average is used in the experiments.

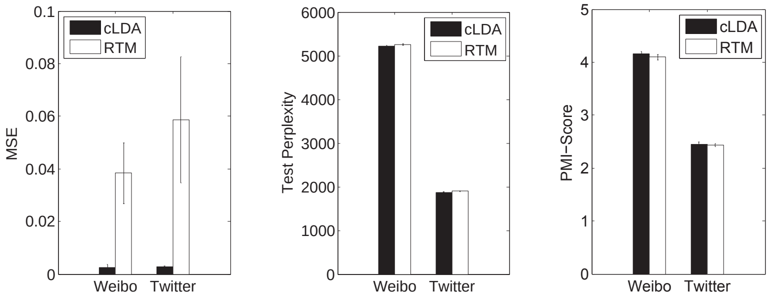
Given p_{ir} and n_{ir}, the algorithms under testing take p_{ir} as the positive feedbacks on (d_i, d_r), and take $n_{ir} = m - p_{ir}$ as number of negative feedbacks. We expect that $\frac{p_{ir}}{m}$ (resp. $\frac{n_{ir}}{m}$) simulates the proportion of similar (resp. different) topics between d_i and d_r.

Metrics We evaluate the effectiveness of our proposed model on predicting human judgements on document similarity (i.e., labels). Given a pair of short document (d_i, d_r) with crowdsourcing labels p_{ir} and n_{ir}, the ratio $\frac{p_{ir}}{p_{ir}+n_{ir}}$ indicates the similarity between the two documents given by the crowdsourcing workers, denoted as s_{ir}. We apply our generative model cLDA to estimate document similarity in terms of the number of positive and negative labels. Specifically, given a set of test document pairs with labels, we compare the predicted document similarity $\widetilde{s}_{i,r} = \frac{\widetilde{p}_{ir}}{\widetilde{p}_{ir}+\widetilde{n}_{ir}}$ against the labeled document similarity $s_{i,r}$. For RTM, we use the predictive link probability as the similarity between documents. We uniformly select 50 test document pairs, and report the *mean squared error* (MSE).

To evaluate the quality of topic modeling, we further report *test perplexity* and *pointwise mutual information* (PMI) score, which are commonly used in topic modeling research as quality measurements. For PMI calculation, we examine the top 10 keywords for each topic, as is used by (Newman et al., 2011). For perplexity calculation, we use the importance sampling method in (Wallach et al., 2009). For each dataset, we randomly select 60,000 documents as the training data and use the rest as the testing data for perplexity and external evaluation archive for PMI. The average results over 10 runs are reported.

Settings For both models, we set the hyperparameter β to 0.1 as in (Griffiths and Steyvers, 2004). To set the optimal number of topics k and hyperparameter α, we conduct a grid search on $\alpha = 0.01, 0.1, 2$ and $k = 5, 10, 15, ...$ using 10-fold cross validation on each training dataset and examine the validation perplexity. For *Twitter* (resp. *Weibo*), k is set to 25 (resp. 40). α is set to 0.1 for both datasets. Notice that even for non-optimal α (e.g., $\alpha = 0.01$) and k (e.g., $k = 10$ for *Twitter*), cLDA is able to achieve improvement in label prediction, test perplexity and PMI. In this section, we only present the results with the optimal settings of α and k. For both models, the burn-in period is set to 1000 iterations.

Results on Comparison of cLDA and RTM In RTM, only observed links are modeled in the inference (Chang and Blei, 2009). Therefore, to compare with RTM, we randomly select document pairs containing same frequent keywords (e.g., tags) with labels $p_{ir} > 0$ and use them as observed links for RTM, as well as labels for cLDA. In this set of experiments, we retrieve 50 such crowd labels.



(a) MSE (b) Test perplexity (c) PMI-Score

Figure 3: Comparison of cLDA and RTM: Average MSE (lower is better), test perplexity (lower is better) and PMI scores (higher is better) on *Weibo* and *Twitter* datasets.

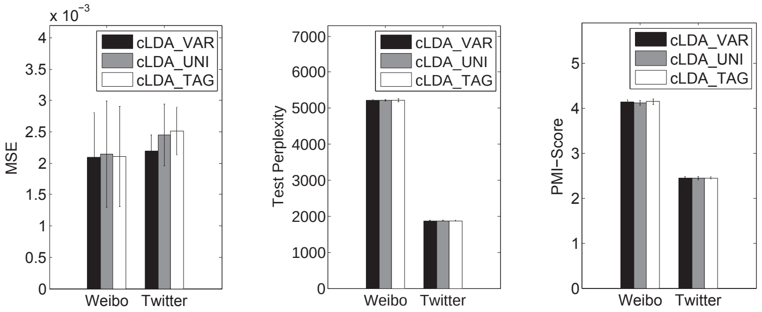
Model	**Topic**
LDA	http 1 bit.li year 2011 2 tinychat shorty 10 part
RTM	song music listen love play sing shorty award np http
cLDA	http bit.li 2011 shorty award January 23 1 nominate vote

Table 1: Top ten words in topics related to *Shorty Awards* generated by different models.

We compare MSE, test perplexity and PMI scores of cLDA and RTM. The results over the two datasets are reported in Figure 3. For document similarity prediction, both models achieve small MSE on both datasets. cLDA outperforms RTM by achieving much lower MSE on both datasets. With 50 crowd labels, cLDA achieves 93% decrease in MSE on *Weibo* dataset. Similarly, the MSE of cLDA is 96% lower than that of RTM on *Twitter* dataset. This is due to RTM's inefficiency in utilizing small amount of link information. Unlike in the original paper of RTM, where it has been applied to document networks with dense links, such small number of links provides insufficient information to train its regression model for the link variables. This further illustrates the advantage of cLDA on crowdsourcing platforms, as it could benefit from small number of crowd labels. For test perplexity and PMI score, both models perform similarly with cLDA slightly outperforming RTM by achieving smaller test perplexity and larger PMI scores on both datasets.

We further investigate the effect of labels/links on the topics generated by comparing top words in topics with and without labels/links. We use *Twitter* dataset as an example. In this dataset, there are several hundreds of tweets about Shorty Awards[2] nominations, which is not a dominant topic in the dataset considering the number of relevant tweets. Table 1 shows the top ten words in topics related to Shorty Awards generated by cLDA and RTM with and without labels/links. As both models become equivalent to LDA without labels/links, we use LDA to denote the results without labels or links. As shown in Table 1, LDA fails to identify this topic with word *award* not appearing in the top ten words. However, both cLDA and RTM capture the co-occurrences of *shorty award*. This is due to information provided by the labels. We notice the labels contain document pairs about Shorty Award and music. For example, there is a document pair selected because of containing common tag #music, one of which is about nominating a singer and songwriter in the Shorty Award and the other is related to universal music group with label $p_{ij} = 1$. This shows both cLDA

[2] https://en.wikipedia.org/wiki/Shorty_Awards



(a) MSE (b) Test perplexity (c) PMI-Score

Figure 4: Comparison of HIT selection methods: Average MSE (lower is better), test perplexity (lower is better) and PMI scores (higher is better) on *Weibo* and *Twitter*.

and RTM can benefit from such label information with increasing ability in identifying topic about Shorty Award. However, RTM tends to mix shorty award with words related to music. This is probably because links are modeled as binary variables in RTM, which makes it difficult to differentiate document pairs with different similarity level. cLDA, on the contrary, can fully utilize label information to better understand topic similarity between documents. In summary, cLDA outperforms RTM in achieving more accuracy in document similarity prediction and higher quality of resulting topic model.

Results on Comparison of HIT Selection Methods For cLDA, we further evaluate the performance of differen HIT selection methods. The results are reported in Figure 4. Notice that document pairs selected by cLDA-TAG are different from those used in the previous experiment, as cLDA-TAG may select document pairs with label $p_{ij} = 0$. cLDA-VAR performs best on document similarity prediction by achieving smallest MSE on both datasets. As its design goal is to minimize the variance of topic distribution $\boldsymbol{\theta}$, which is the key factor in modeling document similarity, it is therefore able to obtain higher accuracy in predicting document similarity than the other two methods. cLDA-TAG performs differently on the two datasets. It performs better than cLDA-UNI on *Weibo* by achieving smaller MSE, but performs worse than cLDA-UNI on *Twitter*. This is probably due to the higher quality of tags in *Weibo*. By examining documents in the two datasets, we found that tags in *Weibo* usually have a stronger correlation with topics than those in *Twitter*. Therefore the document pairs selected by cLDA-TAG and their crowdsourcing labels carry more (accurate) information about topics in documents. This also helps explain why cLDA-TAG obtains the best PMI score on *Weibo*. Except for that, the performance of the three methods are comparable on test perplexity and PMI on both datasets. The results indicate that the choice of which is the best HIT selection method depends on the dataset. cLDA-VAR provides more consistent performance but requires more calculation. cLDA-TAG is more computationally efficient but may be affected by the quality of tags.

7. Conclusion

This paper shows new possibilities of enhancing existing topic modeling techniques with human intelligence by crowdsourcing. We present a new framework to enable crowd workers to contribute their intuitions to topic modeling analysis, by labeling pairs of short texts on their topic similarity. Such intelligence is well absorbed by new inference algorithm to iteratively update the estimations over the latent variable over the revised graphical probabilistic model. Our analysis shows that the algorithm is highly efficient to generate exponentially fast convergence to the optimal solution. Our empirical evaluation reveals that our new method is capable of improving standard LDA algorithm by a significant margin, with only dozens of crowd labels on a much larger text archive.

Acknowledgments

This study is supported by the research grant for the Human Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR).

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, pages 25–32, 2009.

David M. Blei and Jon D. McAuliffe. Supervised topic models. In *NIPS*, 2007.

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

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of ACM*, 57(2), 2010.

Jonathan Chang. lda: Collapsed gibbs sampling methods for topic models. https://cran.r-project.org/web/packages/lda/index.html, 2012.

Jonathan Chang and David M Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. Discriminative relational topic models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(5):973–986, 2015.

David Cohn and Thomas Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. In *NIPS*, pages 430–436, 2001.

Jianguang Du, Jing Jiang, Dandan Song, and Lejian Liao. Topic modeling with document relative similarities. IJCAI, 2015.

Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *PNAS*, 101(suppl 1):5220–5227, 2004.

Ryan Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. Crowdclustering. In *NIPS*, pages 558–566, 2011.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101(Suppl. 1):5228–5235, April 2004.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. Latent topic models for hypertext. In *UAI*, pages 230–239, 2008.

T. Hoffman. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine learning*, 95(3):423–469, 2014.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, pages 897–904, 2008.

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. Topic-link lda: joint models of topic and author community. In *ICML*, pages 665–672. ACM, 2009.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *WWW*, pages 101–110. ACM, 2008.

Ramesh Nallapati and William W Cohen. Link-plsa-lda: A new unsupervised model for topics and influence of blogs. In *ICWSM*, 2008.

Ramesh M Nallapati, Amr Ahmed, Eric P Xing, and William W Cohen. Joint latent topic models for text and citations. In *SIGKDD*, pages 542–550. ACM, 2008.

David Newman, Edwin V. Bonilla, and Wray L. Buntine. Improving topic coherence with regularized topic models. In *NIPS*, pages 496–504, 2011.

Adler J Perotte, Noemie Elhadad, Nicholas Bartlett, and Frank Wood. Hierarchically supervised latent dirichlet allocation. In *NIPS*, pages 2609–2617, 2011.

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

Daniel Ramage, Paul Heymann, Christopher D Manning, and Hector Garcia-Molina. Clustering the tagged web. In *WSDM*, pages 54–63. ACM, 2009b.

Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

Daniel Ramage, Christopher D Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *SIGKDD*, pages 457–465, 2011.

Filipe Rodrigues, Bernardete Ribeiro, Mariana Lourenço, and Francisco Pereira. Learning supervised topic models from crowds. In *HCOMP*, 2015.

Congkai Sun, Bin Gao, Zhenfu Cao, and Hang Li. Htm: A topic model for hypertexts. In *EMNLP*, pages 514–522, 2008.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Kalai. Adaptively learning the crowd kernel. In *ICML*, pages 673–680, 2011.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David M. Mimno. Evaluation methods for topic models. In *ICML*, pages 1105–1112, 2009.

Shuai Wang, Zhiyuan Chen, Geli Fei, Bing Liu, and Sherry Emery. Targeted topic modeling for focused analysis. In *SIGKDD*. ACM, 2016.

Jinfeng Yi, Rong Jin, Anil K. Jain, Shaili Jain, and Tianbao Yang. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *NIPS*, pages 1781–1789, 2012.