One Theme in All Views: Modeling Consensus Topics in Multiple Contexts

Jian Tang*
School of EECS
Peking University
tangjian@net.pku.edu.cn

Ming Zhang
School of EECS
Peking University
mzhang@net.pku.edu.cn

Qiaozhu Mei
School of Information
University of Michigan
qmei@umich.edu

ABSTRACT

New challenges have been presented to classical topic models when applied to social media, as user-generated content suffers from significant problems of data sparseness. A variety of heuristic adjustments to these models have been proposed, many of which are based on the use of context information to improve the performance of topic modeling. Existing contextualized topic models rely on arbitrary manipulation of the model structure, by incorporating various context variables into the generative process of classical topic models in an ad hoc manner. Such manipulations usually result in much more complicated model structures, sophisticated inference procedures, and low generalizability to accommodate arbitrary types or combinations of contexts. In this paper we explore a different direction. We propose a general solution that is able to exploit multiple types of contexts without arbitrary manipulation of the structure of classical topic models. We formulate different types of contexts as multiple views of the partition of the corpus. A co-regularization framework is proposed to let these views collaborate with each other, vote for the consensus topics, and distinguish them from view-specific topics. Experiments with real-world datasets prove that the proposed method is both effective and flexible to handle arbitrary types of contexts.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms

Algorithms, Experimentation

Keywords

Topic modeling; multiple contexts; user-generated content; co-regularization

*This study is done when the first author is visiting University of Michigan.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

1. INTRODUCTION

This is the era when we witness how social media complements, competes with, and eventually substitutes the role of traditional media. Online social communities such as Facebook and Twitter have gained increasing popularity and have eventually transformed into an essential component in our everyday life. Along with this "social movement" is the creation of a huge amount of user-generated content. According to recent statistics, more than 500 million microblogs (i.e., tweets) are posted by Twitter users every day[1]. Such a large volume of user-generated content implies a great opportunity for business providers, advertisers, social observers, as well as data mining researchers. Many interesting data mining tasks have been proposed and performed on the user-generated content in social media, which have not only led to a better understanding of user behaviors in online communities, but also led to more effective techniques of content analysis, information retrieval, and recommender systems.

Textual documents in traditional media, such as newspapers, are professionally formatted and edited, characterized with a benign length of documents and a controlled size of vocabulary. User-generated content in social media, on the other hand, is characterized with extremely short documents, extremely large and evolving vocabularies, and inaccurate uses of language. As an example, a microblog (a.k.a., tweet) in Twitter has a limited length of 140 characters. The sparsity and noise in these user-generated "documents" have introduced new challenges to classical text mining techniques that are effective for traditional media.

One good example is statistical topic modeling [7, 3], which has drawn a lot of attention recently because of its principled mathematical foundation and effectiveness in exploratory content analysis. However, classical topic models usually fail to perform as effectively when applied to user-generated short messages [9]. To improve the performance of topic modeling for social media, a variety of heuristic adjustments have to be applied to the classical models [9, 31].

This is perhaps not surprising given that a topic model essentially works by utilizing the document level word co-occurrences [26, 25]. When the co-occurrence information in a document is sparser and noisier, the performance is inevitably compromised. To effectively apply topic modeling to social media, one has to resort to other types of information beyond word co-occurrences at the document level.

[1]http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html

Luckily, a richer set of context information (e.g., time, location, authorship, friends, followers) is usually observable in social media as a compensation of the compromised quality of the content in individual documents. Indeed, it has been a common practice to utilize various types of contexts in order to improve the quality of topic modeling in social media [22, 31, 16]. Many new topic models have been proposed with one shared intuition – to not only utilize the word co-occurrences at document level but also utilize the signals at the level of various types of contexts. While the intuition of utilizing context information in topic modeling is sound, most of these treatments rely on arbitrary manipulation of the graphical structure of classical topic models. In those contextualized topic models, different types of contexts are integrated into the generative process in an ad hoc manner (e.g., as specific model variables or priors) [24, 1, 18]. This results in models with more sophisticated structures and inevitably more complicated inference procedures [1, 31]. Although these models work well with particular type of context, it is very hard to generalize any of them to handle other types of contexts, or a combination of multiple types of contexts. In real applications, the overhead of finding an effective model structure given a new type of contexts is considerably high.

In this paper, we take the initiative to explore a different direction. Our goal is to find a general solution of utilizing various types of contexts in topic modeling without the overhead of arbitrarily manipulating the graphical structure of the classical topic models. We give a general definition to different types of contexts, interpreting them as multiple views of the "partition of documents" (see Figure 1), which provide different levels of word co-occurrence information. In Figure 1, the collection of text is partitioned by the context of *time* into "pseudo-documents" that are posted at particular time intervals; the context of *author* partitions the corpus into pseudo-documents that are posted by particular users. Under this general definition, the partition of the corpus with original document boundaries is also considered as a special type of context. Among the views defined by different types of contexts, some are more resilient than others, which partition the content into "documents" with a benign length and a more discriminative distribution of topics.

Our intuition is to facilitate the *collaboration* among these views (different types of contexts) in order to improve the quality of the extracted topics. Indeed, a set of topics can be extracted by applying a classical topic model to the corpus partitioned with any of these views. A topic is robust and trustful if it stands out from multiple views (multiple types of contexts) instead of from just one single view. Therefore, if multiple types of contexts can "collaborate" with each other in some way, or if a richer context can "help" a poorer one in some sense, the topics extracted from the corpus are expected to be of a much better quality. This is especially encouraging since most existing contextualized topic models emphasize on the *distinction* between different types of contexts instead of the *collaboration* of them. The distinction of contexts makes the sparse signals in the data even sparser, while the collaboration strengthens the signals. Instead, we encourage different views to collaborate with each other, reinforce each other, and vote for a consensus of topics.

This collaboration process can be achieved through a novel co-regularization framework in which topics extracted from each view (type of context) are regularized by topics from other views. The co-regularization framework simultaneously maximizes the log-likelihood of the collection of documents partitioned with regard to each individual views and meanwhile minimizes the disagreement among different views. An effective algorithm is proposed to optimize the co-regularized objective and the inference procedure remains as simple as those of the classical topic models.

We evaluate the proposed framework with two large-scale real world datasets. Experimental results show that when multiple contexts presented are sufficient for topic modeling, the collaboration of them can effectively improve the quality of topics extracted. The co-regularization framework outperforms other contextualized treatments of topic models, including those with manipulated graphical structures.

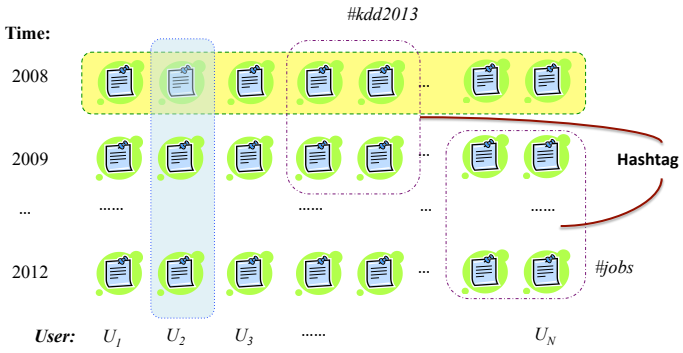


Figure 1: Multiple views of a contextualized collection of user-generated content. Each type of context (time, user, hashtag) defines a partition of the entire corpus. A particular value of a context (e.g., "2008," "U_2," "#kdd2013") defines a pseudo-document (rectangular areas). Original document boundaries define an organic view of the corpus.

2. RELATED WORK

Much existing work has been presented which integrates various types of contexts information into topic models (e.g., [24, 31, 1]). Rosen-Zvi et al. proposed the author-topic model that utilized authorship information for modeling of scientific publications [24]. In their model each author is associated with a multinomial distribution over topics. For each word in a document, an author is uniformly sampled from the set of authors of the paper and then a topic assignment is sampled from the multinomial distribution associated with that author. Ahme et al. developed a model named multi-view topic model that utilizes ideological information [1]. Topics of the corpus are divided into factual topics and ideological-specific topics. For each word in a document, a switching random variable is sampled from a Bernoulli distribution to determine whether the word is generated from a factual topic or an ideological-specific topic, similar to flipping a coin. Many other contextualized topic models are proposed recently, with a common practice to integrate particular types of contexts into the graphical structure of classical models like the latent Dirichlet allocation (LDA). These methods introduced either additional layers to the model (e.g., [24, 14, 10, 30, 29, 8]) or a coin-flipping selection process to select among contexts (e.g., [1, 21]). Although these models work well with given types of contexts,

they all resulted in more complicated model structures and consequentially more sophisticated inference procedures.

Another trend of literature related to our work is topic modeling of user-generated content. Here we present a few representative pieces of work, especially those with a focus on tweets [28, 22, 9]. Weng et al. deployed LDA on the tweets by aggregating all the tweets of the same user and treating each user as a document [28]. This in fact corresponds to the use of the *user* context to partition the data. Hong et al. performed an empirical study of topic modeling in twitter and several aggregation strategies are proposed to train LDA on a dataset of tweets [9]. Work of this kind usually employs a single type of context [31].

These contextualized topic models and treatments are all designed for specific types of contexts and hard to generalize to treat other types of context or a combination of multiple contexts. Our work differs from these by enhancing the collaboration among multiple, arbitrary types of contexts without manipulating the graphical structure of classical topic models. This appears to be a more effective and more generalizable solution of contextualized topic modeling.

Another research direction that is weakly tied to our intuition is *Multi-view Clustering* [2], in which multiple independent views of the data are available and each of them is assumed to be sufficient for clustering. The existing approaches generally aim to exploit the multiple views of the data to discover the clusters that *agree* across the views. Bickel et al. proposed a Co-EM based framework for multi-view clustering [2]. The Co-EM algorithm iteratively performs the E-step in one view, the result of which is passed to an M-step in another view. Kumar et al. proposed a co-training approach for multi-view spectral clustering [12]. Specifically, the spectral embedding from one view is used to constrain the similarity graph used for the other view. In [13], they further proposed a co-regularization framework to regularize the clustering hypotheses across the views. In multi-view clustering, each view corresponds to a representation of the same data points with different features and the target is to cluster the data points by making use of multiple types features. Differently, in our problem a view is provided by a partition of the corpus with a type of context, which reflects the co-occurrence information among words at different levels. The target is to simultaneously assign words and documents into topics by utilizing the different levels of co-occurrence information, which is quite different from the multi-view clustering task.

3. PROBLEM DEFINITION

We start with the intuition that topic modeling essentially relies on the signals of word co-occurrence in documents [26, 25]. Words that frequently co-occur in the same *documents* are likely to be grouped into the same topic, while words do not co-occur tend to be separated into different topics. Such a process echoes the famous assumption of the "*context of situation*" in linguistics, which was first coined by the anthropologist B. Malinowski [23] and then elaborated by J. R. Firth in the quote "*You shall know a word by the company it keeps (J.R. Firth. 1957)* [5]." The basic idea of Firth's perspective is that the **meaning** of a word can be derived from the words with which it co-occurs.

It is a key assumption in all topic models that individual documents present concentrated identities on a few topics. Words in the document are more likely to present the same topic identities as the document. Two connections can be made between topic modeling and the aforementioned linguistic intuition. First, the **meaning** of a word is represented by the identities of the word among a set of **topics**. Second, every **document** provides a **context**, in which the topic identities (meanings) of one word can be derived from the topic identities of the other words co-occurring in this context. This intuitive connection can be further elaborated using a classical topic model, the LDA [3].

3.1 Latent Dirichlet Allocation

The latent Dirichlet allocation (LDA) model assumes that each document is generated from a mixture of topics, with each topic corresponding to a multinomial distribution over all the words in the vocabulary. The detailed generative process of each document $\mathbf{w}$ is described as follows:

1. Sample a document topic proportion $\theta \sim \text{Dirichlet}(\alpha)$.
2. For each word w_n in the document,
   (a) Sample a topic assignment $z_n \sim \text{Multinomial}(\theta)$,
   (b) Sample a word $w_n \sim \text{Multinomial}(\beta_{z_n})$.

α is the Dirichlet prior for the topic proportion. K is the total number of topics and $\beta = \{\beta_k\}_{k=1,\dots,K}$ is the set of topic word distributions.

LDA extracts topics from text by calculating the posterior probability of the hidden variables, the document topic proportion θ and the topic identities of the words $\mathbf{z} = \{z_n\}$, given all observed words $\mathbf{w} = \{w_n\}$ in the documents:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}. \quad (1)$$

This distribution is however computational intractable. In existing literature, variational inference [3] and Gibbs sampling [6] methods have been proposed to approximate the posterior distribution in order to achieve a tractable solution. In either way, such a process essentially attempts to infer the topic identities of every word in a document.

We can clearly observe the critical role of the document-level word co-occurrence. Indeed, from Equation 1 one can observe that the topic identity of an individual word (z_n) not only depends on the word w_n itself, but also depends on all the other words in the same *document* ($\mathbf{w}$). This well elaborates the role of the context of situation (in this case a document). Such co-occurrence signal is carried through to the M-step, where the topic distributions are estimated with all words sharing the same identities of topics. Should the assumption of the "*context of situation*" fail, or should the contexts provide insufficient signals of meaningful word co-occurrences, a topic model is unlikely to perform well.

3.2 Topic Modeling with Multiple Contexts

Topic modeling relies heavily on the contexts of situations that provide sufficient and meaningful signals of word co-occurrences. Naturally, every document provides such a context. A topic model performs well when such "documents" present sufficient signals of word co-occurrences. This condition is true in news articles and scientific papers, where topic models are proved to be effective. This condition is not true in user-generated content, where organic documents are extremely short. Classical topic models (e.g.,

LDA) fail because the document contexts can no longer provide sufficient signals of word co-occurrences.

On the contrary of extremely short documents, a message in social media is usually associated with a rich set of metadata (e.g., time, location, authorship, hashtags, etc.). Some of these metadata variables provide alternative and benign "contexts of situations" rather than the organic documents. In many of these contexts, the assumption that "*semantically related words co-occur*" still holds, and the signals of word co-occurrences become abundant. When the signals are inadequate within organic documents, we can resort to the various types of contexts for help. Motivated by this intuition, we formally define the essential concepts and the problem of multi-contextual topic modeling.

Definition 1. Context, View. *A **context** is an arbitrary subset of the corpus, which corresponds to either an organic document or text that share the same value of a metadata variable. A **type** of context refers to this metadata variable, the values of which defines a partition of the corpus. A **view** of a text collection is defined as this partition of the corpus according to a particular type of context.*

Figure 1 provides an intuitive explanation of contexts and views in a collection of tweets. Tweets are associated with the metadata variables including the *user*, the posting *time*, and the *hashtags* (user created keywords starting with #). A corpus of tweets can then be partitioned by either different *users*, different *time*, or different *hashtags*. Clearly, we have generalized the notion of a "context" so that organic tweets become a special type of context. The organic document boundaries then define a special *view* of the corpus, among with many other views defined by other **types** of contexts (e.g., time, user, hashtags). A specific **context** (e.g., user $= U_2$") becomes a "pseudo-document," which can be derived by aggregating all messages that share the same value of this **type** of context. For example, tweets written by the same user assemble a "pseudo-document" under the *user* **view**; tweets containing the hashtag "#kdd2013" are grouped into a "pseudo-document" under the *hashtag* view.

Definition 2. Topic, Topic Modeling. *A **topic** ϕ is defined as a multinomial distribution over words in the vocabulary V, i.e. $\{p(w|\phi)\}_{w\in V}$. Given a text collection D and a predefined number K, **topic modeling** aims to discover the K salient topics $\{\phi_k\}_{k=1,\cdots K}$ from D.*

In classical topic models, the inference of topics is done with the natural partition of the corpus into *documents*. Now that a *type* of context (or a *view*) also offers a partition of the corpus, one can actually conduct topic modeling with the "pseudo-documents" partitioned using this *view* of corpus instead of using the original documents. By doing this, the topic model utilizes the signals of words co-occurrences at a context level instead of at the document level. A robust topic should appear no matter whether the inference process is done using organic documents or using "pseudo-documents" of a different view. When multiple types of contexts appear, one could imagine a topic model leveraging the signals of word co-occurrences in different views. Topics that appear to be salient in all these views should be the most representative ones of the collection. In other words, these globally salient topics present the *consensus* among multiple types of contexts. Meanwhile, these global topics may present a specific projection onto particular types of context, which represent the *view-specific* interpretations of the consensus topics. In this paper, we aim to address the problem of finding such **consensus topics** and **view-specific topics** by incentivizing individual contexts to collaborate with each other. Formally, we define the problem as follows:

Definition 3. Multi-contextual Topic Modeling. *Let C be the set of context types (views). Each user-generated message m is represented by a pair of vectors $(\mathbf{w}^m, \mathbf{v}^m)$, where $\mathbf{w}^m$ is a vector of words that represent the textual content of the message and $\mathbf{v}^m = (v_1^m, v_2^m, \cdots, v_{|C|}^m)$ represent the context information associated to m. v_i^m presents the observed value (e.g., the date Aug. 8th) of the i-th type of context (e.g., time). $|C|$ represents the total number of different types of contexts. Given a collection of user-generated messages $\{(\mathbf{w}^m, \mathbf{v}^m)\}_{m=1}^M$, **multi-contextual topic modeling** aims to discover the **consensus topics** $\{\phi_k\}_{k=1,2\cdots,K}$ that are the most robust across multiple types of contexts, as well as the **view-specific topics** $\{\phi_{k,c}\}_{k=1,2\cdots,K;c\in C}$ that corresponds to the specific instantiations of the consensus topics in each view.*

Note v_i^m can take either a scalar value or a set of values. For example, a scientific paper can have more than one authors, thus v_i^m may contain a set of names under the *user* context. Similarly, a tweet may contain multiple hashtags.

The problem is substantial challenging because the set of views (C) may be arbitrary. Multi-contextual topic modeling calls for a method that could handle arbitrary types of contexts instead of introducing specific treatments to particular types of contexts. Such a method should facilitate the collaboration among different contexts, with a procedure to "vote" for the consensus topics.

4. MULTI-CONTEXTUAL TOPIC MODELS

In this section, we describe our two methods to tackle the problem. We first introduce a naive treatment with a variation of the classical LDA model. Like many existing contextualized topic models, this model also introduces context variables into the graphical structure. The difference is that it provides a general flexibility to handle arbitrary types of context. Like the existing contextualized topic models, this model is also likely to suffer from the problem of data sparseness. In Section 4.2, based on the motivation to enhance different types of contexts to collaborate with each other, we propose a co-regularization framework that provides a more principled solution to the problem.

4.1 Multi-contextual LDA (mLDA)

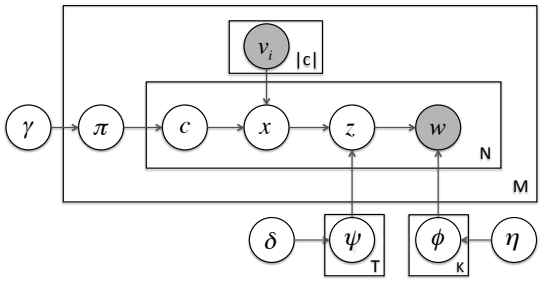


Figure 2: Graphical structure of mLDA.

It is common practice in contextualized topic models, such as the Author-Topic-Model (ATM), to integrate one or more context variables (e.g., the *authorship*) into the generative process or the graphical structure of LDA. Such models are usually capable of handling one or a few given types of contexts (e.g., time, author), but ineffective to incorporate multiple, arbitrary types of contexts. One straightforward solution is to employ a generalization of ATM to multi-contextual topic modeling, by treating all different contexts (e.g., "Aug. 12th," "Chicago") equally as "authors" of a document. Such a treatment fails to distinguish different types of contexts, all of which are flattened into one single view.

Following the common practice but aiming at a generalizable solution, we introduce a multi-contextual LDA (mLDA) model that incorporates multiple, arbitrary context variables as another contextual layer into the graphical structure of LDA. In the corresponding generative process, a particular type of context (i.e., a context variable) is selected with a switching random variable. Specifically, the generative process is defined as follows:

1. For each topic $k \in \{1, \dots, K\}$,
   (a) Sample a multinomial distribution over words $\phi_k \sim \text{Dirichlet}(\cdot|\eta)$;
2. For each context (pseudo-document) x from all views (types of context, partitions),
   (a) Sample a multinomial distribution over topics $\psi_x \sim \text{Dirichlet}(\cdot|\delta)$
3. For each message m,
   (a) Sample a multinomial distribution over context types $\pi \sim \text{Dirichlet}(\cdot|\gamma)$
   (b) For each word w of m
      i. Sample a context type $c \sim \text{Multinomial}(\cdot|\pi)$
      ii. Sample a context value $x \sim \text{Uniform}(\cdot|v_c^m)$, where v_c^m is the set of possible values of context type c in m.
      iii. Sample a topic assignment $z \sim \text{Multinomial}(\cdot|\psi_x)$
      iv. Sample $w \sim \text{Multinomial}(\cdot|\phi_z)$

Gibbs sampling can be used for the inference of the multi-contextual LDA. The conditional probability of the hidden variables for each word w_i is calculated as follows:

$$p(z_i = k, x_i = j, c_i = l | w_i = w, \mathbf{w}_{-i}, \mathbf{z}_{-i}, \mathbf{c}_{-i}, \mathbf{x}_{-i}) \propto (n_{m,l} + \gamma) \cdot \frac{n_{j,k} + \delta}{\sum_{k'}(n_{j,k'} + \delta)} \cdot \frac{n_{k,w} + \eta}{\sum_{w'}(n_{k,w'} + \eta)}, \quad (2)$$

where $c_i = l$, $x_i = j$ and $z_i = k$ represent that the ith word is assigned to context type l, context value j of type l, and topic k. $n_{m,l}$ is the number of times that words in m are generated by context type l; $n_{j,k}$ is the number of times that a word in context j is assigned to topic k; and $n_{k,w}$ is the number of times that word w is assigned to topic k, all excluding the current word token. Note that a user-generated message may contain only a subset of all types of context. For example, not all tweets contain hashtags. In the model, we constraint the value of c_i to the set of observed types of context of the current message. Clearly, when there is only one type of context (e.g., author), the multi-contextual model boils down to a special case that is identical to the author-topic model.

This simple extension of LDA provides a reasonable baseline for mining consensus topics from multiple contexts. The benefit of this model is apparent: it generally handles arbitrary types and combinations of context information, without the need to find a specific manipulation of the model structure for each particular setup. However, we do foresee a potential concern of such a model, or rather a general problem of all existing contextualized topic models that employ a selection process over multiple contexts. That is, the model essentially works to split the words in each message and assign them to different types of contexts. Such a process inevitably makes the available information of a particular context even sparser. This might not be a problem when the document-level information is abundant (e.g., scientific papers), but raises a serious concern for user-generated messages in social media as every message is already very short. We will carefully analyze this issue in Section 5.

The key deficiency of such contextualized topic models (e.g., ATM, mLDA) is that different types of contexts are *competing* for resource (i.e., words), instead of *collaborating* to vote for the consensus topics. Our second attempt is to find a method that facilitates the collaboration among different types of contexts, without making the data sparser for each particular context.

4.2 A Co-Regularization Framework

As mentioned previously, each type of context provides an independent *view* of the corpus, or an independent partition of the content space. Based on each type of context we can derive a collection of "pseudo-documents" and deploy an independent topic modeling process on this collection. Such an approach enjoys the property that information in individual contexts will not be further sparsified, and no arbitrary manipulation is needed on top of the classical topic models. What we need here is a mechanism to push different views to collaborate with each other in order to reach a consensus of the topics. We propose a co-regularization framework to make different types of contexts agree with each other on the topics discovered in each of their own views. Specifically, we adopt a centroid-based regularization schema. We introduce a set of general topic distributions shared across different views (i.e., the consensus topics) and make the topics discovered through each individual view be close to these global consensus topics. In this way, the consensus topics serve as a bridge to make the view-specific topic modeling processes collaborate with each other. Specifically, we introduce the following regularization function to measure the disagreement between the consensus topic distributions and view-specific distributions:

$$R(\beta, \beta^c) = \sum_{k=1}^{K} d(\beta_k, \beta_k^c). \quad (3)$$

$\beta = \{\beta_k\}_{k=1}^K$ is a set of the general topic distributions (consensus topics) and $\beta^c = \{\beta_k^c\}_{k=1}^K$ is the set of topic distributions discovered independently through the view of context type c (view-specific topics). $d(\cdot, \cdot)$ is used to measure the distance between two distributions, and here we adopt the

Kullback−Leibler (KL) divergence [11], i.e.,

$$d(\beta_k, \beta_k^c) = \sum_{v=1}^{V} \beta_{kv} \log \beta_{kv} - \sum_{v=1}^{V} \beta_{kv} \log \beta_{kv}^c \quad (4)$$

Finally, we construct an objective function that consists of the log-likelihood functions of topic modeling in each view penalized by the above regularization term. That is,

$$\Theta = \sum_{c=1}^{|C|} l^c(\alpha^c, \beta^c) - \eta \sum_{c=1}^{|C|} R(\beta, \beta^c), \quad (5)$$

where $l^c(\alpha^c, \beta^c)$ is the log-likelihood of the collection of pseudo-document derived by partitioning the content space using the context type c, i.e.,

$$l^c(\alpha^c, \beta^c) = \sum_{d=1}^{D^c} \log p(\mathbf{w}_d^c | \alpha^c, \beta^c). \quad (6)$$

η is a regularization parameter used to trade-off between maximizing the likelihood of each view-specific topic modeling process and minimizing the disagreement among the topics discovered by each of the processes.

The objective function (5) is still computational intractable as the log-likelihood of each view-specific topic modeling process is intractable. We can still resort to variational inference for a tractable lower bound of the log-likelihoods. Omitting the equivalent details of the variational inference to LDA, we summarize the final updating equations below:

$$\phi_{dnk}^c \propto \beta_{kw_n}^c \exp\{E_q[\log(\theta_{dk}^c)|\gamma_{dk}^c]\} \quad (7)$$

$$\gamma_{dk}^c = \alpha_k^c + \sum_{n=1}^{N} \phi_{dnk}^c \quad (8)$$

$$\beta_{kv}^c \propto \sum_{d=1}^{D^c} \sum_{n=1}^{N} \phi_{dnk}^c w_{dn}^v + \eta \beta_{kv} \quad (9)$$

$$\beta_{kv} \propto (\prod_{c=1}^{|C|} \beta_{kv}^c)^{\frac{1}{|C|}} \quad (10)$$

The updating equations (7) and (8) are the same as the E-step in LDA model [3]. In Equation (9), we can see that the estimation of the view-specific topics depends on not only the identities of all the word tokens in the current view but also the consensus topics across views. In Equation (10), a consensus topic is achieved as the average of the view-specific topics of all views.

We anticipate that this co-regularization treatment would provide a more effective process to mining consensus topics than the multi-contextual LDA and additional benefit of mining view-specific topics. This is because it enhances the collaboration among different types of contexts without making the data sparser. In next section we present experiments using real-world datasets to verify our intuitions.

5. EXPERIMENT

5.1 Datasets

We introduce two real-world datasets for our experiments: one consists of messages (tweets) sampled from Twitter, the leading microblogging site and the other consists of titles sampled from DBLP[2], the online bibliography database .

[2]http://www.informatik.uni-trier.de/~ley/db/

Twitter. We collect a sample of tweets using an official Twitter stream API between October 2nd, 2011 and October 8th, 2011. Through the API, we retrieve tweets of a sample of 2,000 users who have posted at least 100 tweets during the time frame. Stopwords and words that appeared less than 100 times in the whole dataset are removed, which yields a vocabulary of 121,709 unique words[3]. We are able to identify three effective types of contexts in this dataset, including *tweet*, *user*, and *hashtag*. This says, the entire collection can be partitioned into "pseudo-documents" as either individual tweets, the tweets posted by individual users, or tweets containing certain hashtags. The statistics of this dataset are summarized in Table 1.

DBLP. Titles of scientific papers are good instantiations of short textual documents. Metadata information of a paper also provides a rich set of contexts. Such properties connect well with user-generated content, making titles of scientific papers a suitable dataset to verify the effectiveness of multi-contextual topic modeling. We download all the DBLP records that are labeled as conference proceedings. We identify three types of contexts for every record, including the *title* of the paper, the *authors*, and the *conference* (book title) of publication. Words that appeared in less than 5 titles are removed, resulted in a vocabulary of 35,895 unique words. The statistics are summarized in Table 2.

Table 1: Statistics of the Twitter Dataset

Context	# Documents	Avg. doc length by words
User	1,955	355.0
Hashtag	10,769	21.4
Tweet	192,300	3.6

Size of vocabulary: 121,709. A held-out set of 1.9 million tweets is used to evaluate topic coherence.

Table 2: Statistics of the DBLP Dataset

Context	# Documents	Avg. doc length by words
Author	510,097	24.4
Conference	3,804	1095.4
Title	652,521	6.4

Size of vocabulary: 35,895.

5.2 Candidate Models for Comparison

The candidate models are classified into two categories: single-context based and multi-context based.

5.2.1 Single-context Based Methods

- **LDA.** We partition the corpus into "pseudo-documents" according to each single type of context. Then LDA is deployed in each of the partitions. Variational inference is used as the inference algorithm.

5.2.2 Multi-context Based Methods

- **Author-Topic Model (ATM).** The author-topic model was originally proposed to model topics in scientific literature. We adopt a simple generalization which treats each context other than the organic documents as an "author" of the short message (tweet or title), neglecting the different types of context. The Dirichlet priors of topic-word and author-topic distributions in Gibbs sampling are set as 0.01 and 1 respectively.

[3]A few users are dropped who have no nonempty tweet left after this step

- **Multi-contextual LDA (mLDA).** The types of context used are *user* and *hashtag* for the TWITTER dataset, and *author* and *conference* for the DBLP dataset. The hyperparameters of mLDA are set as γ=1, $\delta = 1$ and $\eta = 0.01$.
- **Co-regularization (CR).** We include different combinations of views into comparison. Note that under the co-regularization framework, the organic documents are also treated as an independent view.

In all comparisons, we predefine the number of topics. We did not tune this parameter because our goal is not to find the optimal number of topics but to compare the treatments of contexts. All the results reported below are averaged over 10 independent runs with random initialization.

5.3 Metrics

Finding an objective metric for the comparison of topic models is hard. Many existing studies utilized the perplexity or the likelihood of held-out data. However, such statistics cannot directly measure the quality (e.g., the semantic coherence) of the learned topics. In [4], Chang et al. presented quantitative methods to measure the topical coherence of learned topics. They found that the likelihood of the held-out data is not always a good indicator of topic coherence. To tackle this problem, we introduce two metrics which directly measure the quality of topics discovered.

Topic Coherence. Recently, measuring the semantic coherence of the learned topics has received increasing attention [19, 17]. By measuring how semantic coherent the words in a topic are, one has a better sense of whether the extracted topics are interpretable and to what extent such topics help end-users for exploratory data analysis. In [19], Newman et al. proposed to use the point-wise mutual information (PMI) to measure the semantic coherence of topics. For each topic, the PMI-score calculates the average relatedness of each pair of the words ranked at top-N:

$$\text{PMI-Score}(\mathbf{w}) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad (11)$$

where $\mathbf{w}$ are the top N most probable words of the topic. $p(w_i, w_j)$ is the probability of words w_i and w_j co-occurring in the same document while $p(w_i)$ is probability of word w_i appearing in a document. These probabilities are computed from a much larger corpus. N is set to 20 in our analysis.

Entity Clustering. Another approach to evaluate topic models is to use the learned topics for an external task, and thus assess the quality of topics based on their performance on the task [15, 27]. In this paper, we utilize the learned topics for a clustering task of entities. We select *author clustering* as the external task to evaluate the topics extracted from the DBLP dataset, and *user clustering* to evaluate the topics extracted from the TWITTER dataset. The number of clusters is set the same as the number of topics. Each entity (author or user) is assigned to the topic that is the most prevalent in the pseudo-document corresponding to that entity.

We select the problem of author/user clustering because there are well established metrics for such a task, even when a ground truth is not available. For example, when the interconnections among the authors/users are available, the metric *modularity* [20] is a well accepted metric to evaluate the clusters of the actors (communities) in a network. The modularity score measures the quality of the divisions of a network of actors, which is defined as:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(g_i, g_j). \quad (12)$$

m is the total number of edges in a network. A is an adjacency matrix where A_{ij} is 1 if there exists an edge between node i and node j and 0 otherwise. $\delta(\cdot, \cdot)$ is an indication function, which equals to 1 if node i and node j falls into the same cluster otherwise 0. k_i is the degree of node i.

To calculate the *modularity* of a clustering of authors/users, we first construct a network of the actors. This is straightforward since natural network structures of users/authors exist in both Twitter and DBLP. Specifically, the co-authorship network in DBLP is used to evaluate *author clustering*. If two authors have co-authored at least one paper, then an edge is defined between the two authors. We use the retweet network in Twitter to evaluate *user clustering*. If a user has re-tweeted at least one tweet of another user, then there exists an edge between them. The statistics of the two networks are summarized in Table 3.

Table 3: Statistics of the Networks

Network	# nodes	# edges
Co-author Network	510,097	3,257,571
Re-tweet Network	1,955	1,248

We selected the two metrics, PMI and modularity, because we believe they are direct indicators of the quality of topics rather than the likelihood statistics. A better topic will yield a higher PMI score, and a better set of topics will yield a higher modularity score.

5.4 Experimental Results

We start with a summary of the results of all candidate models on the DBLP dataset. Note that there are three types of context, or three views defined on the DBLP dataset: *author*, *conference*, and *title*. The candidate models produce either the consensus topics or view-specific topics, or both. For single-context based methods, only view-specific topics are generated, when LDA is applied to the view. For ATM and mLDA, only consensus topics are generated. For the co-regularization methods, both view-specific topics and consensus topics are generated. We apply the co-regularization method on different combinations of the views.

The PMI scores of all candidate methods are presented in Table 4. First, we compare the topics discovered based on each single type of context. We can see that LDA based on the *author* view achieves better results than the same method based on the *conference* view, which is far better than the *title* view. This means the performance of LDA on the organic short documents is significantly worse than when applied to "pseudo-documents" partitioned based on either the authors or the conferences. This reassures our observation that when the lengths of documents are too short, they fail to provide sufficient signals of word co-occurrences. Classical topic models like LDA do not work well in such kinds of datasets. Both the *author* context and the *conference* context provide adequate signals of word co-occurrences, where classical topic models perform reasonably well.

Table 4: Topic coherence of topics in the DBLP dataset. Collaboration of multiple views improves both consensus topics and view-specific topics.

Type	Algorithm	Topic PMI			
		Author	Conference	Title	Consensus
Single Context	LDA(Author View)	0.613	–	–	–
	LDA(Conference View)	–	0.569	–	–
	LDA(Title View)	–	–	-0.002	–
Multiple Contexts	ATM (Author-Conference)	–	–	–	0.578
	MLDA	–	–	–	0.577
	CR(Author-Title)	0.622	–	**0.628**	0.612
	CR(Author-Conference)	0.624	**0.641**	–	0.598
	CR(Conference-Title)	–	0.608	0.602	0.606
	CR(Author-Conference-Title)	**0.642**	0.638	0.621	**0.634**

Number of topics is set as 20. The regularization parameter η in the co-regularization framework is empirically set as η=200,000

Table 5: Topic coherence of topics in the Twitter dataset. Collaboration of strong views improves both consensus topics and view-specific topics.

Type	Algorithm	Topic PMI			
		User	Hashtag	Tweet	Consensus
Single Context	LDA(User)	1.94	–	–	–
	LDA(Hashtag)	–	2.54	–	–
	LDA(Tweet)	–	–	-0.617	–
Multiple Contexts	ATM(User-Hashtag)	–	–	–	2.15
	MLDA(User-Hashtag)	–	–	–	2.01
	CR(User-Tweet)	1.82	–	1.52	1.67
	CR(User-Hashtag)	**2.04**	**2.69**	–	**2.32**
	CR(Hashtag-Tweet)	–	2.20	1.36	1.56
	CR(User-Hashtag-Tweet)	1.86	2.50	1.53	1.78

Number of topics is set as 50. The regularization parameter η in the co-regularization framework is empirically set as η=100,000

Let us then look at the PMI scores of multi-contextual topic models. Incorporating multiple types of contexts into the Author-Topic Model and the multi-contextual LDA model does improve over the organic LDA model (LDA with the title view), but the performance is inferior to the best single view of the collection (i.e., LDA with the *author* view). This seems to confirm our concern in Section 4.1 that neither of the two models encourages different types of context to collaborate with each other. The contexts rather compete for resources, which makes the data sparser. As for the co-regularization based methods, we can clearly see that both the view-specific topics and the consensus topics are consistently better than those discovered by single-context based methods and by the two naive multi-contextual methods. This suggests that the co-regularization framework indeed makes different contexts collaborate, not only voting for better consensus topics but also helping each other extract better view-specific topics. Not only did the stronger views (*author*, *conference*) help the weaker view (*title*) significantly, but also did the two strong views reinforce each other. The best consensus topics are achieved when all three views are employed, which is a strong signal of the effectiveness of multi-contextual topics modeling.

Similar findings can be observed on the TWITTER dataset. Remember that there are also three types of context, or three views defined on the TWITTER dataset, namely *tweet*, *user*, and *hashtag*. The PMI scores of the candidate systems are summarized in Table 5. Again, LDA failed to perform well when the single *tweet* view is applied to. Unfortunately the extremely short tweets (up to 140 characters) provided so weak signals for a classical topic model. *Hashtag* seemed to be a rather strong view by its own, significantly outperforming the *user* view, which also performed reasonably well. Among the multi-contextual methods, it is clear that the combinations of the two strong views (*user* and *hashtag*) outperformed all competitors in both view-specific topics and consensus topics. Combining all three views does not improve over the coupling of the two strong views, which should be attributed to the severe sparseness of individual tweets (in average 3.6 words after preprocessing).

Next, we investigate how these topics are useful in particular data mining tasks, by using the view-specific topics for the task of *author clustering* on DBLP and *user clustering* on Twitter. Only the single-context model based on *author/user* view and the co-regularization models that involve the *author/user* view are kept in comparison because only these models output *author/user*-specific topics. The results are summarized in Table 6 and Table 7.

Apparently, view-specific topics using the single view of *author/user* performed reasonably well in the clustering tasks of authors/users. This is not surprising as both the *author* view (in DBLP) and the *user* view (in Twitter) provided sufficient signals for topic modeling, and topic modeling is essentially a way of soft clustering. Interestingly, when combined with another strong view, namely *conference* in DBLP and *hashtag* in Twitter, the *author*-specific topics and *user*-specific topics achieved better performance in author/user clustering. This again confirmed the effectiveness of multi-contextual topic modeling and the co-regularization method.

It is interesting to see that adding the weak view of *title* (in DBLP) and *tweet* (in Twitter) does not further improve the clustering performance. This is consistent with the results using the PMI metric, where the addition of weak views improved the consensus topics but not the view-specific topics of the strong views. The view of organic short documents is simply too weak to provide substantial novel signals for topic modeling as long as stronger views are employed.

Table 6: Author Clustering on DBLP.

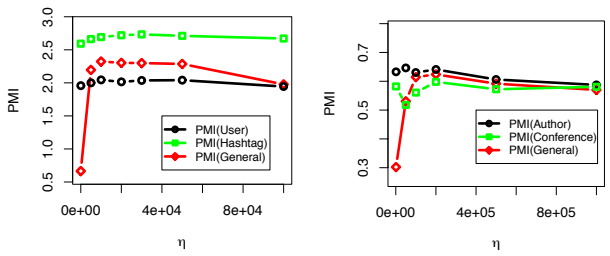
Type	Algorithm	Modularity
Single	LDA(Author)	0.289
Multiple	CR(Author-Title)	0.288
	CR(Author-Conference)	**0.298**
	CR(Author-Conference-Title)	0.295

Table 7: User Clustering on Twitter

Type	Algorithm	Modularity
Single	LDA(User)	0.445
Multiple	CR(User-Hashtag)	**0.491**
	CR(User-Tweet)	0.457
	CR(User-Hashtag-Tweet)	0.480

In summary, the integration of context-based views significantly improves the application of classical topic models on short text documents. When multiple contexts are available, the collaboration of them through the co-regularization framework further improves the consensus topics. The combination of strong views (views that provide sufficient co-occurrence information) also improves the view-specific topics respective to individual type of context. Neither the Author-Topic Model or the multi-contextual LDA model performs as effectively as co-regularization because the two models make the data even sparser through the competition between contexts.

We also investigate the sensitivity of the performance w.r.t the regularization parameter η. In Figure 3, we plot the PMI scores of both the view-specific topics and the consensus topics based on the *user* and *hashtag* views (Twitter) and based on the *author* and *conference* views (DBLP). When η equals 0, the view-specific topics are separately trained w.r.t individual views without any co-regularization. When η becomes larger, the view-specific topics become closer to each other as well as the consensus topics. In general, the performance of the co-regularization framework is not sensitive to the parameter. When the parameter is sufficiently large, the quality of the topics becomes smooth. In practice, the parameter η can be heuristically set as the total number of tokens divided by the number of topics.



(a) Twitter (b) DBLP

Figure 3: Parameter sensitivity w.r.t η.

6. CONCLUSIONS

In this paper, we investigated the problem of exploiting multiple types of contexts for topic modeling in user-generated content. Instead of designing specific manipulations of model structure, we proposed a general co-regularization framework to facilitate the collaboration of different types of contexts. The framework can be easily extended to data sources with arbitrary types and combinations of contexts. Experimental results on two real-world datasets showed that the co-regularization framework successfully incorporated multiple types of contexts, which outperformed contextualized topic models with manipulated graph structures of classical topic models. One interesting problem to be done is how to select the views when many types of contexts are available. Based on our results, the combination of strong contexts significantly outperformed the use of weak contexts. How to measure the strength of a context in topic modeling appears to be a promising future direction.

7. ACKNOWLEDGEMENTS

This work is partially supported by the National Science Foundation under grant numbers IIS-0968489, IIS-1054199, CCF-1048168. It is also partially supported by the National Natural Science Foundation of China (NSFC Grant No. 61272343) and the China Scholarship Council (CSC, No. 2011601194).

8. REFERENCES

[1] A. Ahmed and E. P. Xing. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *EMNLP*, pages 1140–1150, 2010.

[2] S. Bickel and T. Scheffer. Multi-view clustering. In *ICDM*, pages 19–26, 2004.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[4] J. Chang, J. Boyd-Graber, C. Wang, S. Gerrish, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

[5] J. Firth. *A synopsis of linguistic theory, 1930-1955*. 1957.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, New York, NY, USA, 1999. ACM.

[8] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[9] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[10] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *WSDM*, 2011.

[11] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[12] A. Kumar and H. D. III. A co-training approach for multi-view spectral clustering. In *ICML*, 2011.

[13] A. Kumar, P. Rai, and H. D. III. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011.

[14] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM*, pages 375–384, 2009.

[15] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *WWW*, 2008.

[16] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, 2008.

[17] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *EMNLP'11*, pages 262–272, 2011.

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

[19] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *NAACL*, 2010.

[20] M. E. J. Newman. Modularity and community structure in networks. Technical Report physics/0602124, Feb 2006.

[21] M. J. Paul and R. Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*, 2010.

[22] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[23] R. Robins. Malinowski, firth, and the'context of situation'. *Social anthropology and language*, pages 33–46, 1971.

[24] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[25] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, pages 111–120, 2008.

[26] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pages 424–433. ACM, 2006.

[27] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185. ACM, 2006.

[28] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270. ACM, 2010.

[29] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.

[30] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Lpta: A probabilistic model for latent periodic topic analysis. In *ICDM*, pages 904–913, 2011.

[31] 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 *ECIR*, pages 338–349, 2011.