Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering

Ou Jin[*]
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
kingohm@apex.sjtu.edu.cn

Nathan N. Liu
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
nliu@cse.ust.hk

Kai Zhao
NEC Labs China
Beijing, China
zhao_kai@nec.cn

Yong Yu
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
yyu@apex.sjtu.edu.cn

Qiang Yang
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
qyang@cse.ust.hk

ABSTRACT

With the rapid growth of social Web applications such as Twitter and online advertisements, the task of understanding short texts is becoming more and more important. Most traditional text mining techniques are designed to handle long text documents. For short text messages, many of the existing techniques are not effective due to the sparseness of text representations. To understand short messages, we observe that it is often possible to find topically related long texts, which can be utilized as the auxiliary data when mining the target short texts data. In this article, we present a novel approach to cluster short text messages via transfer learning from auxiliary long text data. We show that while some previous works for enhancing short text clustering with related long texts exist, most of them ignore the semantic and topical inconsistencies between the target and auxiliary data and may hurt the clustering performance on the short texts. To accommodate the possible inconsistencies between source and target data, we propose a novel topic model - Dual Latent Dirichlet Allocation (DLDA) model, which jointly learns two sets of topics on short and long texts and couples the topic parameters to cope with the potential inconsistencies between data sets. We demonstrate through large-scale clustering experiments on both advertisements and Twitter data that we can obtain superior performance over several state-of-art techniques for clustering short text documents.

Categories and Subject Descriptors

H.3.3 [**Information Storage and retrieval**]: Information Search and Retrieval—*Clustering*; I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms

Algorithms, Experimentation

Keywords

Short Text, Statistical Generative Models, Unsupervised Learning

1. INTRODUCTION

Short texts play an important role in various emerging Web applications such as online advertisements and microblogging. Sponsored search and display advertisements as commonly found on search result pages or general web pages can usually accommodate just a few keywords or sentences. Similarly, the popular micro-blogging service Twitter restricts the message length to be less than 140 characters. Effective techniques for mining short texts are crucial to these application domains. While many successful text mining techniques have been developed in the past, they have only been designed for and tested on traditional long text corpus such as blogs or newswires. Directly applying these methods on short texts often leads to poor results [12]. Compared with long texts, short text mining has to address two inherent difficulties caused by their highly sparse representations: the lack of sufficient word co-occurrence information and the lack of context information in the text.

To alleviate the sparseness of short texts, a common approach is to conduct "pseudo relevance feedback" in order to enrich the original short text corpus with an additional set of *auxiliary data* consisting of semantically related long texts. This can be achieved by sending the input short texts as queries to a search engine to retrieve a set of most relevant results [18]. Another popular method is to match short texts with topics learned from general knowledge repositories such as Wikipedia or ODP [12, 11]. Once the auxiliary

[*]Part of this work was done while the first author was a visiting student in Hong Kong University of Science Technology.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

data or auxiliary topic is obtained, the data or topics are often directly combined with the original short texts, which are then processed by some traditional text mining models.

While the pseudo relevance feedback based data augmentation approach is a promising approach, one should note that such a process is an inherently noisy operation and there is a risk that certain portion of the auxiliary data may in fact be semantically unrelated to the original short texts. Similarly, the unrelated or noisy auxiliary topics may also bring negative result. Therefore, naively combining the short texts with semantically unrelated long texts or topics could hurt the performance on the short texts. The problem can become even more serious for unsupervised learning on short texts as there is no labeling information to guide the selection of auxiliary data and auxiliary topics.

In this paper, we study the problem of enhancing short text clustering by incorporating auxiliary long texts. We propose a class of topic model - Dual Latent Dirichlet Allocation (DLDA), which jointly learns a set of *target topics* on the short texts and another set of *auxiliary topics* on the long texts while carefully modeling the attribution to each type of topics when generating documents. In particular, we design and compare two mechanisms for DLDA to accommodate domain inconsistencies between the two data sets: (1) using two asymmetric priors on the topic mixing proportions to control the relative importance of different topic classes for generating short and long texts (2) introducing a latent binary switch variable to control whether each document should be generated using target or auxiliary topics. Our DLDA model allows topical structure to be shared in a more flexible manner between the collections of short and long texts and could therefore lead to more robust performance improvement when the auxiliary long texts are only partially related to the input short texts. Clustering experiments on two real world data sets consisting of textual advertisements and twitter messages demonstrated consistent improvements over existing methods for short text clustering, especially when there are many irrelevant documents in the auxiliary data set.

2. RELATED WORK

2.1 Mining Short Text

Due to its importance in popular web applications such as Twitter, short text mining has attracted growing interests in recent years. Hong et al. [4] compared different schemes to train standard topic model on tweets from Twitter. Sriram et al. [19] compared some features for classification to conquer the problem coming with the short of tweets. Mihalcea et al. [8] proposed to measure the similarity of short text snippets by using both corpus-based and knowledge-based measures when acquiring words similarity. Sahami et al. [18] present a novel similarity measure for short text snippets, which utilizes search engines to provide additional context for the given short text, just like query expansion techniques. This similarity measure can also be proven to be a kernel. [21] improved Sahami's work by involving a learning process to make the measure more appropriate for the target corpus. Phan et al. [12, 11] proposed to convert additional knowledge base to topics to improve the representation of the short texts. The knowledge base is crawled with selected seeds from several topics to avoid noise. Hu et al. [5] proposed a three level framework to utilize both the knowledge from Wikipedia and WordNet. Most of these works focus on how to acquire auxiliary data in order to enrich short text. They generally make the implicit assumption that the auxiliary data are semantically related to the input short texts, which is hardly true in practice due to the noisy nature of the pseudo relevance feedback operation.

2.2 Transfer Learning

A closely related area with our work is that of transfer learning [15] which studies how to transfer knowledge from one related auxiliary domain to the target domain to help the learning task on target domain. However, most existing works in this area focused on supervised [20] or semi-supervised setting [14], whereas we need to solve the problem in a totally unsupervised fashion. A similar setting has been considered by Dai et al. [1], who proposed a co-clustering based solution to enhance clustering on target domain data with the help of an auxiliary data set. However, the model is not designed for handling short texts. Moreover, it makes a strong assumption that the word co-clusters are completely shared between the two domains, which, as we will be shown, is much less effective than our much flexible DLDA model.

3. DUAL LATENT DIRICHLET ALLOCATION (DLDA)

Due to its high dimensional yet extremely sparse representations, clustering short texts directly based on the bag of words representation can be very ineffective as we would demonstrate in the experiments. In this section, we develop a topic modeling based approach to discover some low dimensional structure that can more effectively capture the semantic relationships between documents. Directly learning topic models on short text is much harder than on traditional long text. For this reason, [12, 11] proposed to train topic models on a collection long text in the same domain and then make inference on short text to help the learning task on short texts. However, in highly dynamic domains like Twitter where novel topics and trends constantly emerge, it is not always possible to find strongly related long texts via a search engine or a static knowledge base such as Wikipedia. Furthermore, for application domains like advertisement, short texts and long texts are often used for very different purposes. As a result, they may adopt quite distinct language styles. For example, when merchants advertise a product using short banner Ads, the content often emphasizes on the credibility and price aspects. At the same time, in a Web page for selling the product, merchants may focus more on the branding and product features. Similarly, when comparing Twitter messages and Blog articles posted by the same authors, one can also note significant differences in their content as well as language styles.

In the presence of such inconsistencies between short texts and auxiliary long texts, it would be unreasonable to assume that the topical structure of the two domains is completely identical, as done in several previous works [12, 11, 20]. In this section, we describe a better solution to the problem by designing a novel topic model, referred to as the Dual Latent Dirichlet Allocation (DLDA), which can distinguish between consistent and inconsistent topical structures across domains when learning topics from short texts with an additional set of auxiliary long texts. In the following sections,

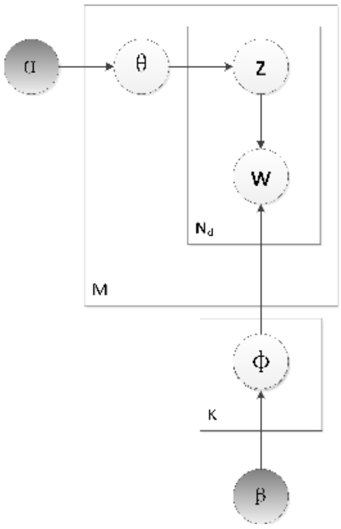

Figure 1: Graphical representation of LDA

Table 1: The generation process of LDA

- For each topic $z \in \{1, \dots, K\}$, draw a multinomial distribution over terms, $\phi_z \sim Dir(\beta)$.
- For each document $d \in \{1, \dots, M\}$
  - Draw a multinomial distribution over topics, $\theta_d \sim Dir(\alpha)$
  - For each word $w_{d,n}$ in document d :
    - draw a topic $z_{d,n} \sim Multinomial(\theta_d)$
    - draw a word $w_{d,n} \sim Multinomial(\phi_{z_{d,n}})$.

we first review the basic LDA model and several related extensions, and then put forward the α-DLDA and γ-DLDA models, which extend the LDA model to cope with domain differences based on two different mechanisms.

3.1 Task Setting and Background

Let $\mathcal{W}^{tar} = \{\vec{w}_m^{tar}\}_{m=1}^{M^{tar}}$ denote the set of short texts from the target domain (i.e., domain of interests), which are to be clustered, and let $\mathcal{W}^{aux} = \{\vec{w}_m^{aux}\}_{m=1}^{M^{aux}}$ denote a set of auxiliary documents consisting of long texts. The auxiliary data set could be extracted from general knowledge base, such as Wikipedia, or extracted from some documents relevant to the target short texts. In this work, we attempt to transfer the topical knowledge from the auxiliary long texts to help with the unsupervised learning task on target short texts. Any long texts that are topically related to the target short texts can be used as auxiliary data.

3.2 Latent Dirichlet Allocation and Extensions

A common approach to dealing with data with high dimensional sparse representation is dimensionality reduction, which has a rich history and literature. A recent trend in dimensionality reduction is the use of probabilistic models. In particular, Latent Dirichlet Allocation (LDA) is a Bayesian probabilistic graphical model, which models each document as a mixture of underlying topics and generates each word from one topic. The generation process of a document is described in Table 1, which corresponds to a graphical model as shown in Figure 1. A document $\vec{w}_d = \{w_{d,n}\}_{n=1}^{N_d}$ is associated a document under a specific multinomial distribution $Mult(\vec{\theta}_d)$ that determines the mixing proportion of different topics within the document. Then, topic assignment for each word is performed by sampling a particular topic $z_{d,n}$ from a multinomial distribution $Mult(\vec{\theta}_d)$. Finally, a particular word $w_{d,n}$ is generated by sampling from a multinomial distribution $Mult(\vec{\varphi}_{d,n})$ over the words in the corpus.

The LDA model is entirely unsupervised, whereas in many applications one cannot expect to have additional knowledge such as class labels, tags, etc. To incorporate such side information so as to arm the LDA model with class labels, several extensions of LDA have been proposed by imposing various constraints on the document-specific topic-mixing proportions $\vec{\theta}_d$. In the DiscLDA model [7], additional supervisory information in the form of document labels is utilized by learning a class-label dependent linear transformation of $\vec{\theta}_d$ with some discriminative criterion.

Similarly, to handle documents annotated with multiple tags such as social bookmarks, the Labeled LDA model [17] learns a topic for each tag and restricts that each document be generated using only those topics corresponding to the set of tags associated with the document. This is achieved by setting a document-specific hyper-parameter vector $\vec{\alpha}_d$. The model was later successfully applied on a large collection of twitter messages annotated with hashtags to map the short messages into different categories [16]. Note that although both [16] and our work attempt to apply topic model for characterizing short text document collections, the focus of our works and their objectives are quite different. [16] studies the utility of hashtags, whereas our work is motivated by the use of additional auxiliary long text documents. In practice, both tags and auxiliary long texts are potential data sources that can be exploited. We believe the two approaches are indeed complementary to each other.

3.3 α-DLDA

Two key ideas are exploited in our work in the DLDA model for coping with domain inconsistencies:

1. We can model two separate sets of topics for auxiliary data and target data. This approach captures the major themes within the two data sets, respectively. Distinguishing auxiliary topics from target topics allows irrelevant or inconsistent topical knowledge in the auxiliary data to be filtered out.

2. We can also use different generative process for auxiliary and target documents, respectively, so that the model favors generating a document using the topics that belongs to its domain.

To realize the first idea, we can simply split the topics in a LDA model into two groups. In particular, we assume that there are K^{tar} topics with parameters $\{\phi_1^{tar}, ..., \phi_{K^{tar}}^{tar}\}$ in the target domain and K^{aux} topics with parameters $\{\phi_1^{aux}, ..., \phi_{K^{aux}}^{aux}\}$ in the auxiliary domain.

To realize the second idea, a simple idea is to properly set the hyper-parameter vector $\vec{\alpha}$, which determines the pri-

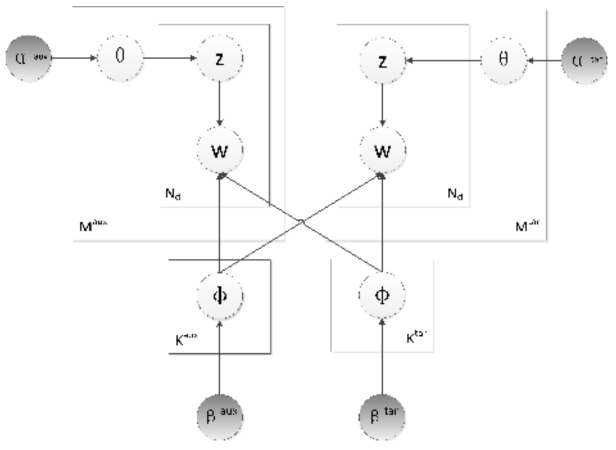

Figure 2: Graphical representation of α-DLDA

or distribution for the document specific mixing proportions $\vec{\theta}$. Traditionally, without any prior knowledge, the entries of $\vec{\alpha}$ are often assumed to be all equal to some small positive value, which corresponds to having no preferences over particular topics. To make the model generate target documents using target topics more often, we can make the entries of $\vec{\alpha}$ correspond to the target topics to be greater than those corresponding to auxiliary topics. The same idea can be applied to designate another asymmetric Dirichlet prior for generating the topic mixing proportions for auxiliary documents. This leads to having two separate asymmetric Dirichlet distributions with parameters

$$\vec{\alpha}^{tar} = [\alpha^{tar}_{aux}, \ldots, \alpha^{tar}_{aux}, \alpha^{tar}_{tar}, \ldots, \alpha^{tar}_{tar}]$$

and

$$\vec{\alpha}^{aux} = [\alpha^{aux}_{aux}, \ldots, \alpha^{aux}_{aux}, \alpha^{aux}_{tar}, \ldots, \alpha^{aux}_{tar}]$$

respectively, which corresponds to the generative process depicted in Figure 2. We refer to this particular variation of DLDA model based on modifying the α parameters as the α-DLDA model. Note that the inference and learning algorithms of the basic LDA model can be easily applied to α-DLDA model, which does not change the model structure but only imposes certain settings of the hyperparameters.

3.4 γ-DLDA

The α-DLDA model uses asymmetric Dirichlet prior to control the relative importance of target versus auxiliary topics when generating a document. However, the asymmetric Dirichlet prior is imposed on all the documents from each domain. In this section, we develop the γ-DLDA model, which constrains that each document be generated using either auxiliary or target topics. This introduces a document-dependent binary-switch variable to be used for choosing between the two types of topics when generating the document. This new mechanism allows the model to automatically capture whether a document should be more related to the target or auxiliary domain.

More specifically, under the γ-DLDA model, each document is associated with (1) a binomial distribution over auxiliary topics versus target topics π_d with Beta prior $\gamma^c = [\gamma^c_{aux}, \gamma^c_{tar}]$ for each corpus $c \in \{aux, tar\}$, and (2) two multinomial distribution over auxiliary topics and target topics $\theta^{aux}, \theta^{tar}$ separately, with a symmetric Dirichlet prior α. The generative process is shown below, and a graphical representation of the model is shown in Figure 3.

Table 2: The generation process of DLDA

- For each target topic $z \in \{1, \ldots, K^{tar}\}$, draw a multinomial distribution over terms, $\phi^{tar}_z \sim Dir(\beta^{tar})$.
- For each auxiliary topic $z \in \{1, \ldots, K^{aux}\}$, draw a multinomial distribution over terms, $\phi^{aux}_z \sim Dir(\beta^{aux})$.
- For each corpus $c \in \{aux, tar\}$
  - For each document $d \in \{1, \ldots, M^c\}$
    - ∗ Draw a multinomial distribution over target topics, $\theta^{tar}_d \sim Dir(\alpha^{tar})$
    - ∗ Draw a multinomial distribution over auxiliary topics, $\theta^{aux}_d \sim Dir(\alpha^{aux})$
    - ∗ Draw a binomial distribution over target topics versus auxiliary topics, $\pi_d \sim Beta(\gamma^c)$
    - ∗ For each word $w_{d,n}$ in document d :
      - · Draw a binary switch $x_{d,n} \sim Binomial(\pi_d)$
      - · if $x_{d,n} = tar$, draw a target topic $z_{d,n} \sim Multinomial(\theta^{tar}_d)$
      - · if $x_{d,n} = aux$, draw a auxiliary topic $z_{d,n} \sim Multinomial(\theta^{aux}_d)$
      - · draw a word $w_{d,n} \sim Multinomial(\phi^{x_{d,n}}_{z_{d,n}})$.

The Beta prior parameterized by γ^c can be used to capture the prior belief on the consistency between the two domains. Here we constrain that $\gamma^{aux}_{aux} > \gamma^{aux}_{tar}$ and $\gamma^{tar}_{tar} > \gamma^{tar}_{aux}$, in order to ensure that the auxiliary topics and target topics focus more modeling the documents in their respective corpus.

3.4.1 *Parameter Estimation with Gibbs Sampling*

From the generative graphical model, we can write the joint distribution of all known and hidden variables given the hyper parameters:

$$p(\vec{w}^c_m, \vec{z}, \vec{x}, \vec{\theta}, \vec{\pi}, \Phi | \vec{\alpha}, \vec{\beta}, \vec{\gamma}^c) \quad (1)$$

$$= \prod_{n=1}^{N_m} p(w^{aux}_{m,n} | \vec{\varphi}_{z^x_{m,n}}) p(x_{m,n} | \vec{\pi}) p(z_{m,n} | \vec{\theta}_m) \quad (2)$$

$$\cdot p(\vec{\theta}_m | \vec{\alpha}) \cdot p(\vec{\pi} | \vec{\gamma}^c) \cdot p(\Phi | \vec{\beta}) \quad (3)$$

The likelihood of a document $\vec{w}$ can be obtained by integrating out $\vec{\theta}, \vec{\pi}, \Phi$ and summing over $z^x_{m,n}$:

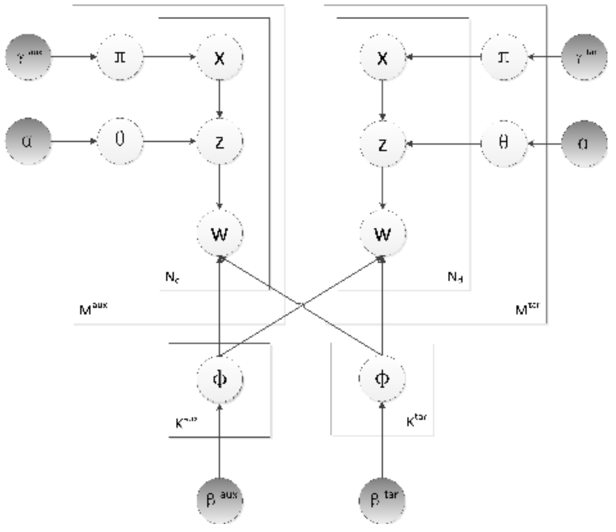

Figure 3: Graphical representation of the proposed γ-DLDA model

$$p(\vec{w}^c|\vec{\alpha},\vec{\beta},\vec{\gamma}^c) = \int\int\int p(\vec{\theta}_m|\vec{\alpha})p(\vec{\pi}_m|\vec{\gamma}^c)p(\Phi|\vec{\beta}) \tag{4}$$

$$\cdot \prod_{n=1}^{N_m} p(w_{m,n}|\vec{\theta}_m,\vec{\pi},\Phi)d\Phi d\vec{\theta}_m d\vec{\pi}_m \tag{5}$$

Finally, the likelihood of the whole data set under the model is $\mathcal{W} = \{\vec{w}_m^{aux}\}_{m=1}^{M^{aux}} \cup \{\vec{w}_m^{tar}\}_{m=1}^{M^{tar}}$ is

$$p(\mathcal{W}|\vec{\alpha},\vec{\beta},\vec{\gamma}) \tag{6}$$

$$= \prod_{m=1}^{M^{aux}} p(\vec{w}_m^{aux}|\vec{\alpha},\vec{\beta},\vec{\gamma}^{aux}) \cdot \prod_{m=1}^{M^{tar}} p(\vec{w}_m^{tar}|\vec{\alpha},\vec{\beta},\vec{\gamma}^{tar}) \tag{7}$$

To estimate the parameters, we need to estimate the latent variables conditioned on the observed variables; i.e. $p(\mathbf{x},\mathbf{z}|\mathbf{w},\alpha,\beta,\gamma)$, where $\mathbf{x}$, $\mathbf{z}$ are vectors of assignments of auxiliary/target topics binary switches and topics for all the words, respectively. We perform approximate inference using Gibbs sampling, a type of Markov Chain Monte Carlo (MCMC) algorithm, with the following updating formulas.

For auxiliary topics $z \in \{1,\ldots,K^{aux}\}$,

$$p(x_i = x, z_i = z|w_i = w, \mathbf{x_{\neg i}}, \mathbf{z_{\neg i}}, \mathbf{w_{\neg i}}, \alpha, \beta, \gamma) \tag{8}$$

$$\propto \frac{n_{w,\neg i}^{aux,z} + \alpha}{\sum_{v=1}^{V}(n_{v,\neg i}^{aux,z} + \alpha_v)} \cdot \frac{n_{d,\neg i}^{aux,z} + \beta}{\sum_{k=1}^{K^{aux}}(n_{k,\neg i}^{aux,z} + \beta_k)} \tag{9}$$

$$\cdot\,(n_{d,x,\neg i}^{aux} + \gamma_x^{c_i}) \tag{10}$$

For target topics $z \in \{1,\ldots,K^{tar}\}$,

$$p(x_i = x, z_i = z|w_i = w, \mathbf{x_{\neg i}}, \mathbf{z_{\neg i}}, \mathbf{w_{\neg i}}, \alpha, \beta, \gamma) \tag{11}$$

$$\propto \frac{n_{w,\neg i}^{tar,z} + \alpha}{\sum_{v=1}^{V} n_{(v,\neg i}^{tar,z} + \alpha_v)} \cdot \frac{n_{d,\neg i}^{tar,z} + \beta}{\sum_{k=1}^{K^{tar}}(n_{k,\neg i}^{aux,z} + \beta_k)} \tag{12}$$

$$\cdot\,(n_{d,x,\neg i}^{tar} + \gamma_x^{c_i}) \tag{13}$$

where $n_{w,\neg i}^{aux,z}$ is the number of times the term w is assigned to an auxiliary topic z. Similarly, $n_{w,\neg i}^{tar,z}$ is the count of w for a target topic z. $n_{d,\neg i}^{aux,z}$ is the number of times a word in document d is assigned to a topic z, while $n_{d,\neg i}^{tar,z}$ is for target topic, $n_{d,x,\neg i}^{aux}$ and $n_{d,x,\neg i}^{tar}$ denote the number of times a word is assigned to auxiliary and to target topics, respectively. c_i denotes the corpus (aux, tar) which the current document is drawn from. For all the counts, subscript $\neg i$ indicates that the i-th word is excluded from the computation.

Table 3: The statistical information of the two corpus

	Short Text		Long Text	
	words/doc	docs	words/doc	docs
ADs	29.06	182209	560.40	99737
TWEETs	9.71	24047	1106.79	4482

After the Gibbs sampler reaches a burn-in state, we can then harvest samples and count the word assignments in order to estimate the parameters:

$$\theta_{d,z}^x \propto n_d^{x,z} + \alpha^x \tag{14}$$

$$\phi_{z,w}^x \propto \theta_w^{x,z} + \beta^x \tag{15}$$

3.5 Clustering with Hidden Topics

After estimating the model parameters and inferring the parameters for new documents, we can represent each document d by θ_d as:

$$f_d = \left[\frac{\theta_{d,1}^{aux}}{S_1^{aux}}, \ldots, \frac{\theta_{d,K^{aux}}^{aux}}{S_{K^{aux}}^{aux}}, \frac{\theta_{d,1}^{tar}}{S_1^{tar}}, \ldots, \frac{\theta_{d,K^{tar}}^{tar}}{S_{K^{tar}}^{tar}}\right]$$

where $S_j^x = \sum_i \theta_{i,j}^x, x \in \{aux, tar\}$. We normalize the scale of each feature in order to reduce the importance of some topics that are overly general, such as a topic that represents the functional words and common language. Such topics tend to occur in most documents, but they lack the discriminative power.

Another important issue is that we should collect sufficient samples of θ in the sampling. Because the texts are short, the result in one sample may vary much, while the average of multi samples will produce a more robust result. This process may be affected by the *label switching* problem caused by the MCMC algorithm. But, in practice, we find that this is not a serious problem.

After having the topic based representations for the short text documents, we can apply the traditional clustering methods on them. Since these features have already involved the knowledge from the auxiliary data, clustering on the new representation may achieve better results, as we will demonstrate in the experiments.

4. EXPERIMENTAL RESULTS

4.1 Data Set

In order to evaluate our algorithm DLDA, we conduct experiments on two real data sets, a collection of advertisements (ADs) from an online consumer to consumer (C2C) shopping Web site and a collection of tweets (TWEETs) from Twitter.

The short texts in the ADs data set are a collection of text-based display advertisements crawled from an e-commerce Web site of a commercial company. To obtain the auxiliary long texts, we randomly crawled some product web pages listed on the e-commerce site's home page. Each advertisement has been labeled with 42 classes according to the product taxonomy used by the site. To create the TWEETs data set, we crawled 197,535 tweets from 405 Twitter users' time line. We then filter out 24,047 tweets (12.1%) with hashtags (i.e. semantic annotations added by the author of a tweet). In these tweets, there are 4,282 tweets (17.8%) containing URLs, we therefore crawled the content of the referenced URLs to form the set of auxiliary long texts. Some statistics of these two data sets are shown in the Table 3, from which we can see that the short texts in each data set contain a very small fraction of the number of words in the corresponding long texts. Moreover, the short texts in the TWEETs data set contain very few words on average and are much shorter in length when compared with the ADs data set.

We performed standard data preprocessing including stopword removal on the raw text. For the TWEETs data, we filter out the non-semantic symbols such as mentions of user names (@user), shorted URLs, etc. We also extract hashtags (#tag) from the text, since we rely on the hashtags as the semantic labeling information for evaluating the clustering quality.

It should be noted that our model does not require any correspondence structures between the short texts and the long texts, such as those indicating which tweet contains which URL. This type of correspondence is also not available at all in the ADs data set. The short and long texts in both data sets are of very different nature and are only topically related. The crawled Web pages in the ADs data set may not necessarily contain any information related to the products mentioned in the advertisements. Similarly, in the TWEET's data set, the referenced URLs are often news articles and blogs written by the users themselves, whereas the tweet messages are mostly the users' subjective comments. Thus, the tweets and the web pages are very likely about the same topics, though the style of the languages can be quite different.

4.2 Evaluation Criteria

In the experiments on the ADs data, we used *Entropy*, *Purity* and *Normalized Mutual Information (NMI)* as the evaluation criterion. Purity is a simple and transparent evaluation measure, which can help us understand the quality of the clustering result directly. Entropy and NMI can be information-theoretically interpreted. NMI allows us to trade off the quality of the clustering against the number of clusters, which the Purity measure cannot achieve. The larger the Purity and NMI values are, the higher the quality of the clustering is. Similarly, a smaller entropy means better performance.

Due to the lack of ground truth of the TWEET's data, we use the hashtags contained in the tweets as their hidden-meaning indicator. Since a lot of tweets are assigned more than one hashtags, we propose to employ *Davies-Bouldin Validity Index* (DBI) [2] calculated on the hashtags as the evaluation criterion. DBI is a function of the ratio of the sum of within-cluster scatter to between-cluster separation, which is defined as:

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \left\{ \frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)} \right\}$$

where n is the number of clusters, Q_x is the centroid of cluster x, $S_n(Q_x)$ is the average distance from all elements in cluster x to centroid Q_x, and $S(Q_i, Q_j)$ is the distance between centroids c_i and c_j. Here we use the Euclidian distance as the distance measure for both functions $S_n(\cdot)$ and $S(\cdot, \cdot)$. Since the algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies-Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest DBI is considered the best algorithm based on this criterion.

4.3 Baseline Methods and Implementation Details

We compare our DLDA with the following methods:

Direct Direct clustering method without incorporating auxiliary data. We use CLUTO [1] with the TF-IDF representation.

LDA-short Learning a LDA from short texts and directly clustering with the θ learned.

LDA-long Learning a LDA from long texts and then apply to short text, which is proposed in [12]. We use GibbsLDA++ [2], the implementation mentioned in the paper.

LDA-both Learning a LDA from the combined collection of long texts and short texts as proposed in [20].

Self-Taught Clustering (STC) This is a state of the art unsupervised transfer learning method [1], which exploits an additional set of unlabeled auxiliary data for clustering target data. It does co-clustering [3] on both data sets simultaneously and uses shared word-clusters to bridge the two domains.

For LDA-short, LDA-long, LDA-both and our DLDA model, we first build the topic based representations of the short texts following the procedures described in Section 3.3, and then use CLUTO to obtain the document clusters.

For each application of the LDA-based algorithm, we repeat the step for five times and compute mean. We tune each algorithm to its best parameter setting. For the experiments on the ADs data, we set the cluster number to 42, since the data set is labeled with 42 classes. For the experiments on the TWEET's data, we set the cluster number to 50.

4.4 Result

The experimental results on two corpuses are shown in Table 4 and Table 5, respectively. As expected, our methods α-DLDA and γ-DLDA clearly outperformed all the other baseline methods on both data sets. Due to the high dimensionality and sparseness of the representations in short texts, directly clustering the short texts gave the poorest performance. LDA-short outperformed Direct on both data sets, which demonstrates the benefit of using topic based low dimensional representation for the short texts.

[1] CLUTO:`http://glaros.dtc.umn.edu/gkhome/views/cluto/`

[2] GibbsLDA++:`http://gibbslda.sourceforge.net/`

[3] Co-clustering:`http://www.cse.ust.hk/TL/index.html`

Table 4: The performance of all the evaluation methods on ADs

	Purity	Entropy	NMI
Direct	0.280	0.681	0.215
LDA-short	0.305	0.650	0.250
LDA-long	0.360	0.599	0.310
LDA-both	0.362	0.594	0.315
STC	0.320	0.636	0.259
α-DLDA	0.383	0.578	0.336
γ-DLDA	**0.392**	**0.553**	**0.364**

Table 5: The performance in DBI of all the evaluation methods on TWEETs

	DBI
Direct	18.270
LDA-short	15.002
LDA-long	14.249
LDA-both	13.960
STC	13.342
α-DLDA	12.485
γ-DLDA	**11.748**

All methods that utilized auxiliary long texts can be observed to significantly outperform both the Direct and LDA-short, which clearly shows the value of transferring knowledge from auxiliary long texts. Interestingly, the LDA-long model, which ignores all the short texts, performs better than LDA-short. We believe that this is because learning topic models on short texts is inherently much more difficult.

Both variations of the proposed DLDA model consistently beat the LDA-both and STC algorithms. LDA-both does not distinguish the target domain from the auxiliary domain, and can suffer from domain inconsistencies as a result. The STC model distinguishes target and auxiliary domains, but lacks a document-level mechanism for determining if a document is more related to the target or the auxiliary domain. The empirical results confirm that the DLDA can effectively address these difficulties.

Finally, when comparing α-DLDA with γ-DLDA, we see that α-DLDA treats the documents in each set uniformly. However, for γ-DLDA, the document-dependent binary-switch variables may only help select those auxiliary long texts that are related to the target domain. This document level mechanism gives γ-DLDA an extra flexibility over α-DLDA, which can lead to the observed superior performance.

4.5 Influence of the Parameters and Auxiliary Data

We conduct additional experiments to show the influence of the parameters and auxiliary data. These experiments are done on the ADs corpus with NMI as the evaluation metric. Due the stochastic nature of the algorithm, we repeat each algorithm for 5 times and report the mean as well as standard deviation values.

Hyperparameters.

In [3] parameters are set as $\alpha = 50/K$ and $\beta = 0.01$. In our work, we set $\gamma_{tar}^{aux} = \gamma_{aux}^{tar} = \gamma_{small}$ and $\gamma_{tar}^{tar} = \gamma_{aux}^{aux} = \gamma_{big}$, where $\gamma_{small} < \gamma_{big}$. Since the model is not very sensitive to this prior, we use $\gamma_{small} = 0.2$ and $\gamma_{big} = 0.5$ in the experiments.

For α-DLDA, we rely on the setting of α prior to constrain the topic selection. We have four parameters α_{aux}^{tar}, α_{tar}^{tar}, α_{aux}^{aux}, α_{tar}^{aux} to be tuned. We set $\alpha_{aux}^{aux} = 50/K$ and vary the other three in $\{0, 0.01, 0.05, 50/K\}$.

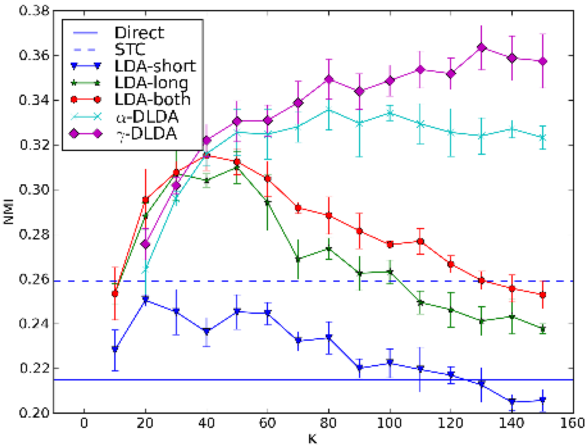

Figure 4: The performance of all the algorithms on ADs with different K

Topic Numbers.

The performance of all the LDA-based algorithms on different topic numbers is shown in the Figure 4 and 5 ,for both the ADs and TWEETs data sets. For α-DLDA and γ-DLDA, $K = K^{aux} + K^{tar}$ represents the overall complexity of the model. With the same K, the computational complexities of these LDA-based methods are nearly the same.

From these results, we can find that the method without any additional information is poor, while the other algorithms achieve much better performance. Those topic models with small-sized topics have similar performance, but perform much differently when the number of topics is large. The method that treats auxiliary data and target data differently can handle a larger number of topics with better result.

We also examine different K^{aux} for γ-DLDA. The result is shown in the Figure 6. For a set of K values ranging between 60 and 140 , we plot the model performances as K^{aux} goes from 10 to K. The value of K controls the overall complexity of the topic model, whereas tuning K^{aux} allows us to enforce different degrees of topics shared among the target and auxiliary data. Given a fixed K, reducing K^{aux} is equivalent to enforcing more shared topics between the two domains, whereas increasing K^{aux} allows more auxiliary domain-dependent information to be captured by the auxiliary topics. From Figure 6, we can clearly see that the optimal performance is achieved with neither very small nor very large K^{aux} values, which shows the important tradeoff between topic-sharing and domain inconsistencies.

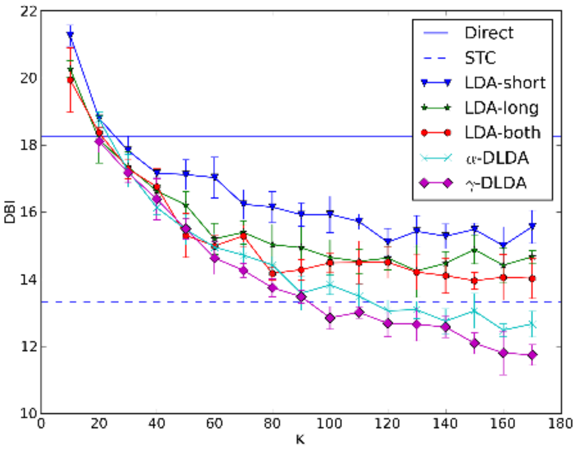

Figure 5: The performance of all the algorithms on TWEETs with different K

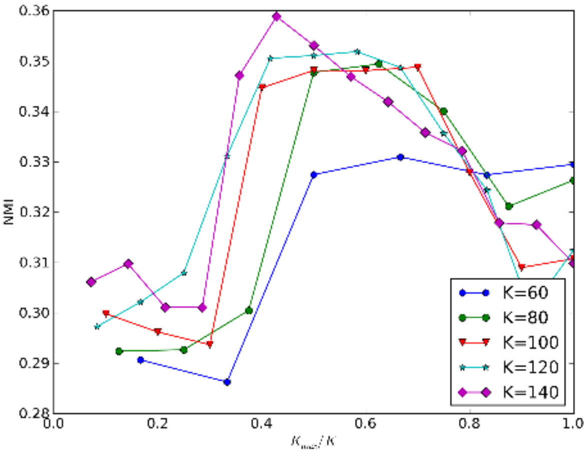

Figure 6: The influence of different K^{aux} and $K^{aux} + K^{tar}$ for γ-DLDA

Auxiliary Data.

The amount of auxiliary data involved in the learning process also influences the result. In the second set of experiments, we examine such an influence by varying the amount of auxiliary data used for training from 10% to 100%. The result is shown in the Figure 7. We can see that even with a small amount of auxiliary data, the performances of α-DLDA and γ-DLDA are already much better than the other methods. Furthermore, increasing the amount of auxiliary data can lead to consistent improvement. With more flexibility, γ-DLDA begins to perform better when more auxiliary data is involved.

Irrelevant Data.

We also compare the robustness of our algorithms α-DLDA and γ-DLDA with the other two topic model based methods: LDA-long and LDA-both. In this experiment, the auxiliary data is mixed with some documents that are randomly selected from a general knowledge base. There are certainly a lot of useful documents in the general knowledge base that share topics with our target documents. But without any filtering, much more irrelevant data can be included. These documents are not strongly related to the target data and can be considered as noise in the auxiliary data. We control the noise level by inserting different number of irrelevant documents into the auxiliary data. The performance of different models under different noise levels is shown in the Figure 8. The results of LDA-long and LDA-both are not very stable when more irrelevant documents are added to the auxiliary data whereas our method γ-DLDA achieved stable performances with different amounts of noise. This proves that, by explicitly considering domain inconsistency, we can effectively improve the robustness of short text clustering with auxiliary long texts.

When dealing with the auxiliary documents without noise, there is no huge difference between γ-DLDA and α-DLDA. In Figure 8, we found that simply treating the documents uniformly will suffer a lot of trouble. Manually adjusting the α priors can lighten this problem, but it requires much human effort. Without carefully tuning the α priors, the performance can drop dramatically.

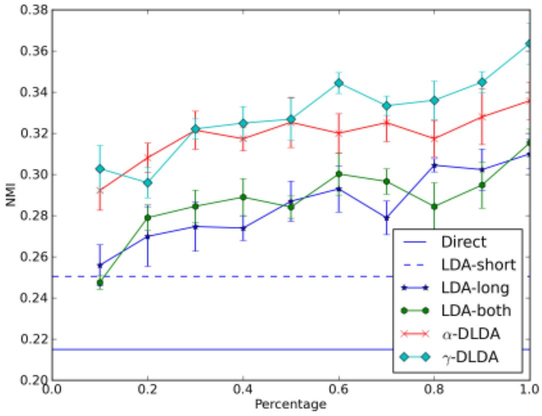

Figure 7: The influence of the number of auxiliary data used

The result of γ-DLDA can also be explained by the following analysis. We calculate the π when mixing the irrelevant data. The π is calculated by

$$\pi^c_{d,x} \propto n^c_{d,x} + \gamma^c_x.$$

Then the average value of π for auxiliary documents and target documents are calculated separately, as shown in Figure 9. With more irrelevant data, the model automatically adjusts the relevance between target documents and auxiliary topics. This adaptability helps the model escape from using many irrelevant topics.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a type of novel topic model for enhancing short text clustering by incorporating auxil-

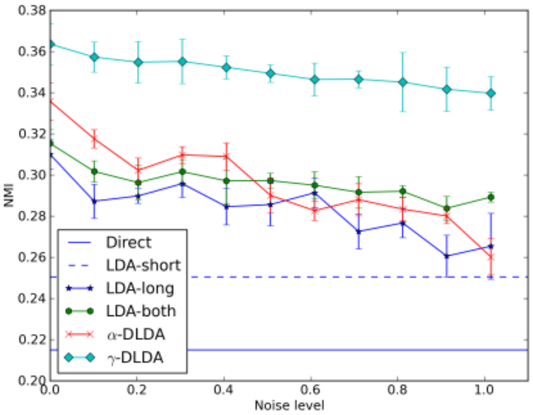

Figure 8: The influence of irrelevant data in auxiliary data

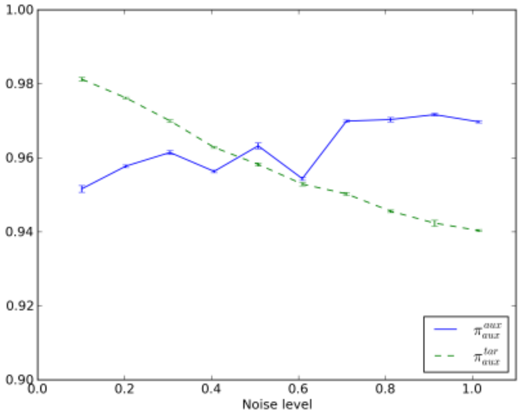

Figure 9: The variation of π with irrelevant data in auxiliary data

iary long texts. The model jointly learns two sets of latent topics on short and long texts. When considering the difference between the topics of the auxiliary and target data, our method can robustly improve the result of clustering on target data, even when there are a lot of irrelevant documents in auxiliary data. The experimental result shows that DLDA can outperform many state-of-the-art methods. This helps validate that, by considering the difference between auxiliary data and target data, the clustering quality on short text can be improved.

In the future, we wish to evaluate other forms of domain difference criteria between data sets, and evaluate their performance when knowledge transfer is conducted between the data. We will also consider other forms of topic models to enable more effective learning.

6. ACKNOWLEDGEMENT

We thank the support of Hong Kong RGC project 621010, and Projects of International Cooperation and Exchanges National Natural Science Foundation of China 60931160445.

7. REFERENCES

[1] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 200–207. ACM, 2008.

[2] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, April 1979.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, April 2004.

[4] L. Hong and B. Davison. Empirical study of topic modeling in twitter. *1st Workshop on Social Media Analytics*, 2010.

[5] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 919–928. ACM, 2009.

[6] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 389–396. ACM, 2009.

[7] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2008.

[8] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, pages 775–780. AAAI Press, 2006.

[9] C.-T. Nguyen, X.-H. Phan, S. Horiguchi, T.-T. Nguyen, and Q.-T. Ha. Web search clustering and labeling with hidden topics. *ACM Transactions on Asian Language Information Processing*, 8:12:1–12:40, August 2009.

[10] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, October 2010.

[11] X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi, and Q.-T. Ha. A hidden topic-based framework towards building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering*, 99, 2010.

[12] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 91–100. ACM, 2008.

[13] X. Quan, G. Liu, Z. Lu, X. Ni, and L. Wenyin. Short text similarity based on probabilistic topics. *Knowledge and Information Systems*, 25:473–491, December 2010.

[14] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international*

conference on Machine learning, ICML '07, pages 759–766. ACM, 2007.

[15] R. Raina, A. Y. Ng, and D. Koller. Constructing informative priors using transfer learning. In *ICML '06*, pages 713–720, New York, NY, USA, 2006. ACM.

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

[17] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[18] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 377–386. ACM, 2006.

[19] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 841–842. ACM, 2010.

[20] G.-R. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 627–634. ACM, 2008.

[21] W.-T. Yih and C. Meek. Improving similarity measures for short segments of text. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pages 1489–1494. AAAI Press, 2007.